8-2014

Y-Chromosome Polymorphism in United Arab Emirates

Safa Yaqoub Yousif Alkhayyat AL Hammadi

UAEU

United Arab Emirates University

College of Science

Biology Department

Y-CHROMOSOME POLYMORPHISM IN UNITED ARAB EMIRATES

Safa Yaqoub Yousif Alkhayyat Alhammadi

This thesis is submitted in partial fulfillment of the requirements for the Degree of Master of Science in Environmental Sciences

Under the direction of Dr. Khaled Amiri, Dr. Ahmed Al Marzouqi and Dr. Abdulmajeed Al Khajeh.

August 2014

DECLARATION OF ORIGINAL WORK

I, Safa Yaqoub Yousif Alkhayyat Alhammadi, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of the thesis titled as "Y-chromosome polymorphism in United Arab Emirates", hereby I declare that this thesis is an original work done and prepared by me under the guidance of Dr. Khaled Amiri, Dr. Abdulmajeed Al Khajeh, in the College of Science at UAEU and Dr. Ahmed Al Marzouqi in Abu Dhabi general head quarter. This work has not previously formed as the basis for the award of any degree, diploma or similar title at this or any other university in the UAE. The materials borrowed from the other sources and included in my thesis have been properly acknowledged.

Student's Signature..... **Date**27/08/2014.

Approved by

Thesis Examining Committee:

1) Advisor (Committee Chair): Dr. Khaled Amiri

Department of : Biology

College of : Science

Signature : _____ Date: 27/8/2014

2) Member: Dr. Sofyan Alyan

Department of : Biology

College of : Science

Signature : _____ Date: 27/8/2014

3) Member (External Examiner): Dr. Mohsin Sulaiman Alaamri

Division of : DNA fingerprinting & forensic

Institution: Advanced Biotechnology Centre

Signature : _____ Date: 27/8/2014

Accepted by

Master's Program Coordinator:

Signature ..... Date 10/9/2014

Dean of the College:

Signature ..... Date 10/9/2014

Copy .....of .....

ABSTRACT

Human Y chromosome is a specific male marker and it consists of the largest non-recombining segment in human genome that is the hallmark of Y chromosome population-based studies.

This study establishes an extensive Y polymorphism profile of the UAE population and to our knowledge, this is the largest study carried out in the UAE. DNA samples were genotyped for 17 polymorphic STR from 345 unrelated Emirati male.

In UAE population, the analysis of the allele frequency clearly shows that each locus has a predominant allele. It is also apparent that alleles for most loci are clustered over a narrow range where approximately 60% - 80% of the population is sharing a specific allele for the locus.

Moreover, the highest diversity were observed at locus DYS458 = 0.9, DYS385-A = 0.9 and DYS385-B = 0.9. Therefore, they should be considered as the most variable and most informative markers for forensic testing. While, loci with the lower diversity are the least informative loci (i.e. DYS392 which equal 0.437).

The UAE population is largely heterogeneous and a total of 301 different haplotypes were identified. There are 271 unique haplotypes and 22 haplotypes were shared between two individuals. There are three cases where four, five and six individuals are sharing identical haplotype. Moreover, there are three different haplotypes shared by five individuals. This is likely due to the sharing of most common recent ancestors. This brings the discrimination capacity to approximately 90% and haplotype diversity 99.885%. This is fundamental to understanding the

degree of heterogeneity in UAE population and can reflect the pattern of the migration, geographic influence, and cultural influences. Secondly, the study provide an array of haplotype that can serve as database for forensic uses. To this end, United Arab Emirates population is diverse and are genetically close to neighboring countries. Analysis of molecular variance (AMOVA) show no significant genetic differences within UAE population or population residing the Arabian Gulf region.

Dedication

I loving dedicate my thesis to my beloved parents, sisters and brothers who have never failed to give me a moral support and for giving me all I need and encouraging me to continue my higher education.

Moreover, I dedicate this thesis to the person who is my manager, friend and brother Ibrahim Al-Hosani who has really been there through the hard times.

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

1.1 Literatures review:

As the 1920s began, the existence of the Y chromosome was debated and Theophilus Painter published a very short article claiming the presence of Y chromosome in humans and other primates. Y chromosome, unlike mitochondrial genome, is inherited paternally and approximately 95% percent of Y chromosome is inherited as one haplotype (Rosa et al., 2007). Therefore, Y chromosome represents an invaluable record of all mutations that have taken place along male lineages throughout the evolution. The importance of Y chromosome is not limited to determining the sex of a species, but involves the differential gene dosage, and shows a characteristic pattern of inheritance in males and females. Y chromosome constitutes approximately 2% of the human genome and consists of highly polymorphic loci. Therefore, Y chromosome analysis is used in many applications spanning different fields and includes paternity determination, human migration, archeogenetics, and paleontology (Rosa et al., 2007). Another important application of Y chromosome analysis is in forensic science it is used to solve murder case, rape, and even exoneration of convicted felons (Carvalho-Silva et al., 2001).

There are a flurry of studies involving Y chromosome and human population such as studies in human migration, influence of geographical location, and shaping of human culture (Gaetano et al., 2009). The beauty of Y chromosome studies lends itself to several slowly mutating alleles on non-recombining Y chromosome (NRY) that can be used to group human Y chromosomes into vertical paternal lineages called Y chromosome haplotype. Many more studies are available that address the

importance of Y chromosome studies (Jobling et al., 2003). The Y chromosome study is also used to determine disease association in a population (Sezgin et al., 2010). These studies provide important indicators of human susceptibility to diseases and response to drugs (Krausz et al., 2004). Despite the importance of Y chromosome studies, there is a dearth of information on the population of Arabian Gulf regions. The aim of this study is to characterize Y chromosome specific loci that can decipher diversity or uniqueness of the UAE population structure. For a comprehensive survey of Y chromosome polymorphism, it is important to study a larger population across the UAE. This will decipher patrilineage genetic diversity in the UAE population which is one of the important elements to elucidate cultural influence, the origin of surnames, dialect, disease association, population structure, and establishment of forensic databases.

1.2 Human Genome:

Human genome is diploid and contains two copies of chromosomes, each inherited from a parent. During gametogenesis, the genome is reduced, to haploid state that comprises of 3 x 10^9bp (Krawczak et al., 1994). Genetic information (except for mitochondrial genome) resides in the nucleus of the cell and is organized into physical structures called chromosomes whose structural state changes during cell division. In general, Chromosomes contain the genetic information of cells called Deoxyribonucleic Acid (DNA), and it controls many cellular functions (Daniel, 2008). Chromosomes are normally transmitted as an intact unit from parent to children. According to the random assortment principle, these markers residing on one chromosome are inherited together and they exhibit genetic linkage. In contrast,

markers on different chromosomes are generally inherited independently of one another and they do not demonstrate linkage (Stryer, 1999). Conversely, markers that show genetic linkage, implying that these are close together on the same chromosome. Linked loci are transmitted in clusters or haplotypes unless recombination processes changes their phasing (Backer et al., 1995). That is, they are associated together more often than chance would predict.

Human genome is distributed over 46 chromosome and 23 homologous pairs (Daniel, 2008). Twenty-two pairs of these chromosomes are called autosomes and labeled according to length, longest to shortest, and one pair is called sex chromosomes, X and Y. A normal person receives two copies of chromosomes, one set is derived paternally and one set maternally. The diploid cells results from the fusion of 2 haploid cells sperm (23 chromosomes) and egg (23 chromosomes) to have an embryo or gametes of 46 chromosomes. Y chromosome determines the sex of the embryo. Males have one X and one Y chromosome (XY), while females have two X chromosomes (XX). Therefore, the gender of the embryo is determined by the paternal contribution and DNA testing in the laboratory can determine this. Human genome are very similar to one another and on average two individual share 99.7% of the sequence of their DNA. The remaining 0.3% (~10 million nucleotides) bears the variation that exists among individuals (Tishkoff et al., 2004). Some extranuclear DNA is also present in the mitochondria that are located in the cytoplasm of the cells, and can be used for human identification. Although the majority of these nucleotide differences are neutral.

1.3 The structure and composition of human genome:

In 1869, Frederich Miescher a German biochemist was the first to observe DNA and unfortunately, researchers did not realize the importance of this molecule for a long time. In 1953 James Watson, an American biologist, and Francis Crick, a British physicist, proposed a model for the structure of the DNA molecule. This model was based on research by Rosalind Franklin, Maurice Wilkins, and other scientists (Daniel, 2008). Their works opened the doors to a new field of research and one of these fields is known as molecular genetics, which is the study of the function-structure relationship and the study of inheritance and variation at molecular level. At the most basic level, DNA is found in nearly all-living cells except RNA viruses. DNA is a molecule that is shaped like a twisted ladder and it consists of two separate strands that intertwine to form a double helix, a structure that resembles a spiral staircase as shown in figure 1. Each strand of DNA is composed of a series of smaller molecules called nucleotides. In turn, each nucleotide is itself made up of three smaller molecules or called primary components: a nitrogenous base, monosaccharide deoxyribose, and a phosphate group. The nucleotides are joined together by covalent bonds to form polynucleotide chains. There are four different DNA nucleotides, each defined by a specific nitrogenous base: Adenine abbreviated as (A), Thymine abbreviated as (T), Guanine abbreviated as (G), and Cytosine abbreviated as (C). DNA contains two types of nitrogen-containing bases. Adenine and Guanine, whose ring-shaped molecules have six members, are called purines; while Cytosine and Thymine with fused five- and six-member rings are called pyrimidines. According to complementary base pairing concept, Adenine and

Thymine always base pair with each other by two hydrogen bonds while, guanine and cytosine are attached together by three hydrogen bonds see figure 1 (Daniel, 2008).

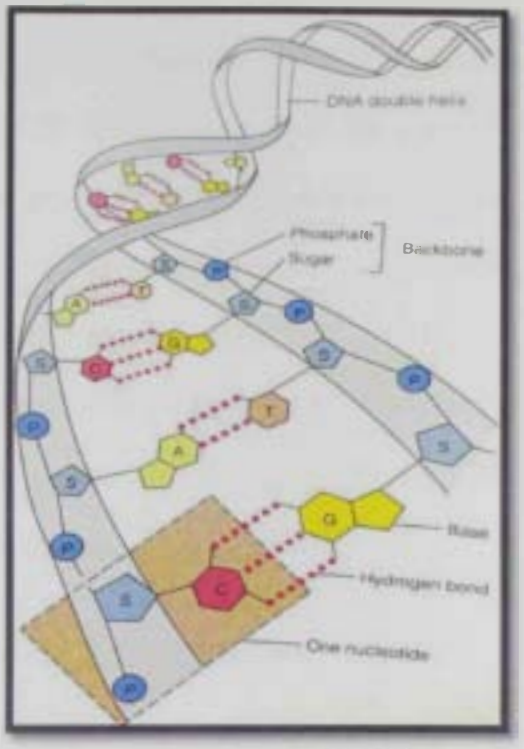


Figure 1: Component of the DNA structure

Naturally, complementary base-pairing is responsible for the ability to accurately replicate DNA molecules, with its genetic information and pass it on to the next generation (Daniel, 2008).

1.4 Human Y-Chromosome:

The human Y chromosome is one of the fastest evolving parts of the human genome. In the last decade, Y chromosome has been increasingly used to investigate the migrations, evolution, and range expansions of modern humans and it also has attracted a great deal of attention owing to its supremacy in male sex determination and unique haplotype status in the genome. In 1905 during a study of the mealworm,Tenebrio molitor at Bryn Mawr College, Nettie Stevens has identified Y

Chromosome as a sex-determining chromosome, and named it Y because it always existed in pairs with X chromosome. Which was discovered in 1890 by Hermann Henking (Haley et al., 2010). There are several interesting attributes and biological features of the human Y chromosome; one of these features is carrying a limited number of functional genes with a high proportion of repeat elements (Ali et al., 2003). Additionally, it has important biological male-specific functions with direct consequences on male fitness including male fertility and testis determination. Features, like pattern of inheritance among others, make the study of Y chromosome polymorphisms very useful for inferences of population histories, forensic applications and paternity analysis (Ali et al., 2003).

1.5 Structure of the Y-Chromosome:

Y Chromosome is one of the smallest chromosomes in the human genome (Hsrries et al., 1986). It represents only 2% of the human genome in males and it contains about 60 million base pairs (Sher et al., 2002). Y Chromosome can be divided into two small tips known as pseudoautosomal region (PAR) and non-pseudoautosomal region (male-specific) that are many times larger than PAR. There are two regions in the pseudoautosomal region called PAR1 which is located at the tip of the short arm (Yp) with approximately 2.5 million base pairs in length and PAR2 which is less than 1 million base pairs in length and it is located at the long arm (Yq) of the Y chromosome (Ali et al., 2003; see figure 2).

On the other hand, the rest of the Y chromosome (about 95%) which is termed the non-recombining portion (NRY) or male-specific region (MSY) does not undergo sexual recombination during meiosis. It is always in a haploid state, and therefore, is

transmitted intact through paternal lineages. The NRY remains the same from the father to son unless a mutation occurs (Butler, 2011).

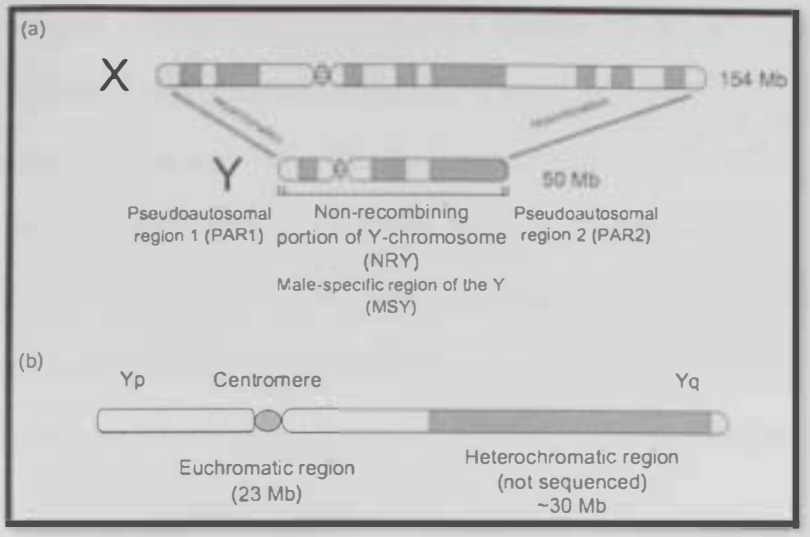


Figure 2: Schematic of X and Y sex chromosomes.

The Y chromosome is composed of both euchromatic and heterochromatic regions of which only the 23 Mb of euchromatin has been sequenced (Butler, 2011). The Y chromosome sequence composition, like any other chromosome, consists of unique sequence (normally represent coding sequence), and repetitive. The repetition are either repeated randemly or interspersed and generally they are referred to as satellite and its subset depending on the length of unit repeat. The following sections aim at introducing satellites DNA characteristics and its subset (for review see Jobling et al., 2003).

1.6 Satellites DNA:

There are four major classes of tandem DNA repeats that have been classified within the human genome: satellite 1, satellite 2, satellite 3 and satellite 4 (Milkos et al., 1997). Around 20% of the human genome is composed of various satellite DNA families (Ballantyne et al., 1989). Satellite DNAs were originally identified by the separate banding of part of the genomic DNA in equilibrium density gradient centrifugation. Satellite DNA is classified according to their genomic localization, such as centromeric, telomeric or dispersed along the chromosome. Satellite DNA is primly located near the centromeric regions of the chromosomes and the repeat unit length can extend for several thousand base pairs (Willared et al., 1987). Alphoid class is another major tandem repeat DNA species, which is composed of approximately 2% of the whole genome (Manuelidis, 1987). The Alphoid satellite consists of tandem arrays of approximately 171 bp monomers (Yang et al., 1982), and it is localized to the pericentromeric region of each human chromosome (Jobling et al., 1998).

1.7 Minisatellite DNA:

Minisatellite term refers to the Variable Number Tandem Repeat (VNTR) that is a class of highly repetitive satellite DNA. By definition, minisatellite means the section of DNA that consists of a short series (10–60 base pairs) of nucleotide. These occur at more than 1,000 locations in the human genome. There are two types of minisatellite; the first is called "Hypervariable Minisatellite" that have a core unit 9–24bp long and are found mainly at the centromeric regions, and the second is called "Telomeric Minisatellite" that have core units 6 bp long, and have thousands of

repeated sequences at the telomeres. During the last two decades, a number of highly variable regions in the human genome have been detected and characterized. The use of VNTR polymorphisms, subsequently, has become one of the most successful changes in the field of forensic medicine for personal identification and paternity testing. In Y chromosome two minisatellites have been described. The first was MSY1 (DYF155S1) and which has a 1% mutation rate per generation. According to Jobling, this marker was reported as the most variable locus in Y chromosome (Jobling et al., 1996). While the other minisatellite described is MSY2 (DYS440) contains only two units of three or four copies of 99-110bp repeating unit (Gill et al., 1995). One of the main advantages of these markers is that they are highly polymorphic in some lineages.

1.8 Microsatellites DNA:

Microsatellite is defined as Simple Sequence Repeats (SSRs) or also Short Tandem Repeats (STRs). It is another source of polymorphisms of the human genome, which belong to the family of repetitive non-coding DNA sequence. They can be found in coding and non-coding regions. Microsatellite is characterized by the length variation in tandem arrays of simple repeat sequences of 2-6 base pair. STRs provide a rich source of polymorphic markers resulting from variations in the number of copies of the repeated pattern. They are similar to VNTR loci and minisatellite loci but the latter contain longer repeat units. STR loci display several advantages that make them attractive as genetic markers. They are very plentiful, averaging one trinucleotide tandem repeat locus for every 15kb in the human genome, and they are amenable to PCR amplification by using flanking sequence primers. The resulting

amplification by fragment of the individual STR loci ranges from 100 to 400bp. The characterization of a large number of highly polymorphic STR loci along with the construction of well-defined allelic ladders for several of the most easily interpreted loci, allows for an increased use of these systems in forensic analysis and paternity determination. The Y chromosome is very rich in several classes of repeated DNA sequences. Using these microsatellite loci is a very useful tool in the forensic identification of male DNA in rape cases with male and female fractions (see sections below; Butler, 2011).

1.9 Short Tandem Repeats (STRs):

The human genome is full of repeated DNA sequences. STR is one example of these short repeated sequences. The nature of STR give rise to highly polymorphism region. These tandem repeats clusters are characterized by blocks of DNA of some common sequence which is repeated over and over in tandem. These repeated sequences come in various sizes and are classified according to the length of the core repeat units. Short tandem repeats are found around the chromosomal centromere and consist of a short repeat unit ranging approximately 2 to 6 base pairs in length (Butler, 2011). The number of repeats in STR markers can be highly variable among individuals, which make these STRs effective tools for studying polymorphism and genetic variation between individuals and populations. At the beginning of 1996, the FBI Laboratory launched a nationwide forensic science effort to establish core STR loci for inclusion within the national database known as CODIS (Combined DNA Index System). The 13 CODIS loci are: CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11.

These loci are nationally and internationally recognized as the standard for human identification (Collins et al., 2004). However, the number of loci that are used for genetic variation as the knowledge of the polymorphic loci is increasing.

1.9.1 Types of STR loci:

Microsatellites can be classified based on size, the nature of the repeated unit or their position within the genome. The microsatellite, based on its repeat uniformity, can be clustered into four categories as shown in table 1 (Urquhart et al., 1994). Often the repeats are interrupted with few base pairs.

1.9.2 Simple STR loci:

Generally it contains one repeating unit with equal length and sequence and they are not interrupted by other sequences. Examples of this type are D5S818, D13S317, D7S820, D16S539, TPOX and CSF1PO loci. The simple STR can be further subdivided into simple one repeating sequences such as HUMFES/FPS and simple with non-consensus alleles like HUMTHO1.

1.9.3 Compound STR loci:

This type of STR loci consists of two or more adjacent simple repeat sequences (HUMGABRB15) or consists of a compound with Non-consensus alleles, such as (HUMvWFA31/A).

1.9.4 Complex STR loci:

This type of STR loci allele commonly consists of regular tetranucleotide repeat units with interspersion of dimer, trimer and hexamer invariants.

1.9.5 Hypervariable STR loci:

This locus is highly polymorphic and contains complex compound regions that can show many alleles that differ by one base pair. Examples of this type are HUMACTBP2 or SE33 and D11S554 loci.

Table 1: Different types of STR loci

No	Type	Loci	Repeat Examples
1	Simple	D5S818	$(AGAT)_{7-16}$
2	Compound	HUMvWA	$(ATCT)_2(GTCT)_{3-4}(ATCT)_{9-13}$
3	Complex	D21S11	$(TCTA)_{4-6}(TCTG)_{5-6}[(TCTA)_3TA(TCTA)_3$ $TCA(TCTA)_2TCCA](TCTA)_{8-16}(TATCTA)_{0-1}TC$
4	Hypervariable	ACTBP2	A common repeat structure (AAAG) with different mono, di, tri, tetra and hexamer invariants that are scattered throughout the locus.

1.10 Y-STR Markers:

There are two categories of DNA markers, which are used to examine Y chromosome diversity. The first marker is called bi-allelic loci, which exhibits two possible alleles. Bi-allelic markers are also referred to as unique event polymorphisms (UEPs) and that is due to their low mutation rates, which is estimated from 10^{-8} to 10^{-9} per generation. The example of this marker includes single nucleotide polymorphisms (Y-SNP) and Alu element insertion (YAP). The second marker called multi-allelic loci, which includes two minisatellites, and several hundred STR markers. Their results are characterized as haplotypes and it can be used to differentiate Y chromosome haplotypes with high resolution due to their higher mutation rates (Butler, 2011).

Y chromosome DNA analysis can be performed with either Y-STRs which defines haplotype or Y-SNPs which define haplogroup. Y-STRs results exhibit more variability due to the rapid change in mutation of Y-STRs (mutation rate ≈1 in 10^3 compared to Y-SNPs (mutation rate ≈1 in 10^9) (Butler, 2011). Thus, Y-STRs have a greater use in forensic identification of male DNA such as in rape cases and paternity determination in deficiency cases where the alleged father is missing. The majority of Y-STRs results in a single polymorphic fragment upon PCR amplification, but there are some Y-STRs which originate from regions that are duplicated on the Y chromosome, resulting in the presence of two amplicons of variable size (i.e.DYS385, DYS459 and recently DYS464 (Butler, 2011). The following figure 3 illustrates the relative positions of 17 markers STRs loci commonly used in human identification setting.

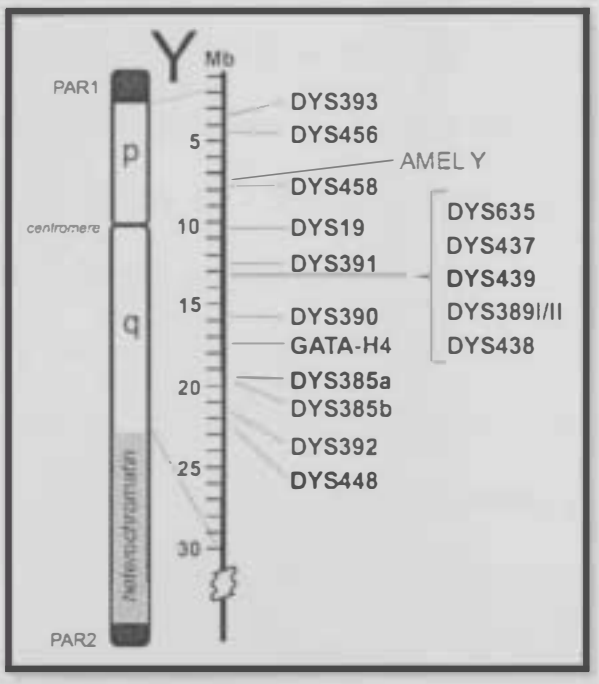


Figure 3: Relative positions of 17 markers STRs loci

1.11 Background of United Arab Emirates (UAE):

Figure 4: Map of the United Arab Emirates.

The United Arab Emirates (UAE) is one part of the Gulf Cooperation Council (GCC) which consists of six gulf countries including Bahrain, Kuwait, Qatar, Saudi Arabia, and Oman. In addition, the United Arab Emirates is located in the eastern part of the Arabian Peninsula, extends along part of the Gulf of Oman and the southern coast of the Arabian Gulf. The United Arab Emirates is a constitutional federation of seven emirates: Abu Dhabi, Dubai, Sharjah, Ajman, Umm al-Qaiwain, Ras al-Khaimah and Fujairah. The federation was formally established on the 2 of December 1971. The UAE occupies an area of 71023.6 Sq. km, along the south-eastern tip of the Arabian Peninsula. As shown in figure 4, Qatar lies to the west, Saudi Arabia to the south and west, and Oman to the north and east. The UAE lies between ° east (UAE statistical center, 2010). Furthermore, the capital and the largest city of the federation

in the UAE is Abu Dhabi, which accounts for 87 percent of the UAE's total area. However, Ajman is the smallest emirate, encompassing only 259 square kilometers.

According to several studies, researchers discovered that the United Arab Emirates has a long history, recent findings on the eastern side of the Hajar Mountains and in the western region of Abu Dhabi having pushed the earliest evidence of Man in the Emirates back by hundreds of thousands of years. The arrival of envoys from the Prophet Muhammad (PBUH) in 630 AD heralded the conversion of the region to Islam. By 637 AD Islamic armies were using Julfar (Ras al-Khaimah Emirate) as a staging post for the conquest of Iran. Over many centuries, Julfar became a wealthy port and pearling center from which great wooden dhows ranged far and wide across the Indian Ocean.

The UAE population can be divided into three sectors Urban (hadhar), nomadic (bedu) and rural. Bedouin people are traditionally inhabitants of the Arabian Gulf who claim descent from two male lineage: Adnani and Qahtani (Nature, 2010). In addition, these nomadic people are called people of the desert, who are animal owners and move about with their camels, sheep and goats in search of grazing and concentrate around their wells. Some of them used to gather firewood from the scrub and bring it into the coastal towns. However, Hadhar people made their living from the sea. Moreover, they lived by fishing and pearl industry. In contrast, rural people used agriculture as a main way of life in the eastern mountain area and oases. However, what was overwhelmingly rural and bedu a generation ago is now preponderantly urban (Al-Sayegh, 1998). Therefore, in this study the UAE population was divided into three regions Northern, Eastern and Western based on the proximities.

In this study we are addressing the heterogeneity of the UAE population with regard to paternal linage inheritance. The UAE population as per our knowledge has not been genetically characterized on a large scale. This study is one step in a series of studies of the UAE population. The specific aim of this study is to define the UAE population structure and hierarchy through analysis of 17 STR of Y chromosome.

1.12 Objectives:

Several objectives of this study can be defined as the following:

1- To establish a thorough profile of Y polymorphism in the UAE.

2- To study the pattern of migration and geographic influence.

3- To evaluate the analysis of paternal lineages for forensic purpose.

4- To investigate the variance of 17 short tandem repeat (STR) loci in the UAE population.

5- To study haplotype and haplotype frequency in the UAE population through the analysis of 17 polymorphic short tandem repeat (STR) loci.

CHAPTER II: METHODS AND MATERIAL

2.1 Samples Collection:

Buccal swap samples were collected from 723 native males from the United Arab Emirates population of which 345 were analyzed for this study. This is the largest study in the region. The samples were obtained randomly from unrelated individuals encompassing the UAE population (Abu-Dhabi, Dubai, Sharjah, Fujurah, Ras Al kaimah, Ajman, Um Alqawaen, Khalba and Khorfakkan). The random sample collection complies with the regulations of many scientific communities such as the International Society for Forensic Genetics (ISFG) and the National Research Council (NRC). According to these bodies all samples deposited in the international databases should be collected randomly in order to eliminate biases.

2.2 Consent Form:

The buccal samples were collected with written consent from the subjects. The consent Form was approved by Al-Ain Medical District Human Research Ethical Committee which is an accredited organization of Federal Wide Assurance (FWA) and compliant with ICH/GCP Standards. The consent form was written in English and translated into Arabic. The subjects were informed of the nature of study before the relevant information, such as name, family name, place of origin, was obtained.

2.3 Buccal Swabs Packaging:

Samples collected from outside of Al-Ain city were transferred in an icebox to the National DNA Database (NDNAD) Laboratory. On arrival, consent forms were separated from the sample and the samples were stored at -20°C for further use.

2.4 Preparations and Sample Batching:

Once the samples were received in the lab, a unique barcode was assigned to each sample. Subsequently, the samples were batched together by robotic process called TubeStar® (manufactured by Peak Analysis & Automation, UK) and placed in an extraction rack for processing. Briefly, TubeStar® is a positional logging station that allows batch DNA extraction simultaneously. The samples are distributed in an extraction rack that consist of 96 wells of which 86 wells are for DNA samples and 10 wells are for controls (positive and negative extraction controls, blank tubes for Allelic Ladders and PCR controls).

2.5 Storage Condition:

DNA samples were analyzed by AmpFlSTR Y filer kit (Life Technology). The kit is stored at -20^{O}C and includes AmpFlSTR Y filer Primer set, AmpliTaq Gold® DNA polymerase, AmpFlSTR Y filer PCR Reaction Mix, AmpFlSTR Control DNA 007 and AmpFlSTR Y filer Allelic Ladder. On delivery to the lab, AmpFlSTR Y filer Allelic Ladder was isolated and stored at 4^{o}C for post amplification as to prevent cross contamination. All other reagents were stored at 2-8^{O}C. Chemicals that contain fluorescent (i.e. AmpFlSTR Y filer Allelic Ladder and Pico Green etc.) are sensitive to light and were stored in a dark area.

2.6 General Sample Procedures:

Figure 5: Analysis workflow for all samples.

The workflow in figure 5 shows the general sample processing. The DNA for each batch was extracted with silica membrane method in NDNAD Laboratory with Biorobet Universal Instrument; then the batch is moved for quantification with Pico Green fluorescence on Infinite pre 200 Instrument platform. After quantification of DNA, the DNA (96-well Quant DNA®) is amplified on Tetrad PCR machine. Finally, the amplified DNA plate was loaded on 3500 XL Genetic Analyzer which (24 capillary capacity) for fragment resolution.

2.6. 1 DNA Extraction Process:

All the DNA extracted processes were performed according to the manufacture's protocol. The first step of analyzing the genomic DNA is to isolate DNA from tissues or cells using a combination of physical and chemical methods. Many different platforms and technologies are available. One of the platforms is called Silica Membrane Matrix. Biorobet Universal System from Qiagen® is an instrument that uses silica membrane matrix method. It is designed to perform fully automated medium for high throughput applications in 96-well format. DNA can be purified from swabs, blood, and forensic samples etc. Moreover, it provides a rapid and efficient method for DNA extraction from nucleated cells. The basic principle of silica matrix can be described in four stages. The first stage is to lyse the tissue by breaking up the cells with reagents like ATL® buffer, physical agitation in the presence of proteinase K. ATL® buffer contains sodium dodecyl sulfate (SDS) detergent that disrupts cell membrane and dissociates protein DNA complex. Proteinase K solution, meanwhile, degrades proteins including DNA scaffolding proteins and other protein debris. The second stage is to isolate DNA from the cell by

adding AL® lysis buffer and Ethanol. AL® buffer is used to lyse the cell and inactivates nucleases. It contains guanidinium hydrochloride chaotropic salt which removes water from hydrates molecules in a solution and renders DNA susceptible to binding to silica matrix in the spin plate. As DNA is insoluble in alcohol, ethanol is used to precipitate DNA out of fluid suspension. Therefore, it increases DNA affinity to bind to silica spin column for further elution of DNA through the spin plate. The third stage is DNA Purification. The purification step uses wash buffer (AW1 and AW2). Each of the buffers has a different ethanol concentration. AW1 being the more concentrated buffer followed by AW2. This ensures the DNA remains bound to the silica matrix on the column. The buffers act to dissolve and remove cellular debris and contaminants that are not bound to the silica matrix. The vacuum draws these through the spin column. The result of this stage is to leave clean and purified DNA bound to a silica matrix. AW® buffer contains Tris[tris(hydroxymethyl) aminomethane,(HOCH2)3CNH2] and Ethylenediaminetetraacetic acid (EDTA). EDTA removes debris by chelation metal ions (attaching molecules to itself using the ionic charge). The final stage of DNA extraction is elution. In this step Nuclease free water (NFW) is used. Water was added to change the binding conditions through re-hydrating the DNA and removing hydrogen bonds, and subsequently eluting the DNA from the column. The last step was performed at 60° C and pH 8.3 for a maximum efficiency.

2.6. 2 Quantification Process:

DNA quantification is critical for molecular analysiswhich involves STR amplification. There are several methods used to establish the concentration of the DNA in solution. The most common method of DNA quantification is spectrophotometric quantification. However, there are several other methods to spectrophotometric quantification including Fluorometry, AluQuant, Quantitative PCR (qPCR) and UV fluorescence in presence of a DNA dye. In this study, DNA was quantified by using DNA-binding dye, namely, PicoGreen® (Life Technology). One of the advantages of this method is its ability to quantify a small amount of double stranded DNA as little as 25 pg/ml of dsDNA in the presence of ssDNA. The assay is linear over three orders of magnitude and has little sequence-dependence, which allows it to accurately measure DNA from many sources, including genomic DNA, viral DNA, miniprep DNA, or PCR amplification products. DNA-dye complex is then subjected to light at 480nm DNA/PicoGreen which emits light at 520nm. The light emitted correlates with the concentration of the DNA. Standard DNA concentration curve is generated with Sonicated Human placenta DNA (SHP). Fluorescence plate reader software. PicoGreen® registers emission of the light. Fluorescence material are light sensitive and working solutions were prepared freshly each time by adding 50µl of Pico Green® and 10 ml of Nuclease-free water and immediately wrapped in aluminum foil.

2.6. 3 Amplification Process:

2.6.3.1 DNA Amplification using AmpFLSTR® Yfiler® PCR kit:

This step involves specific amplification of fluorescently tagged DNA. AmpFLSTR® Yfiler® PCR Amplification Kit (Life Technologies) was used to co-amplify 17 Short tandem repeat (STR) loci plus a sex determining (Amelogenin) locus. These loci are: DYS456, DYS3891, DYS390, DYS3891I, DYS458, DYS19, DYS385a/b, DYS393, DYS391, DYS439, DYS635, DYS392, GATA-H4, DYS437, DYS438 and DYS448. The AmpFLSTR® Yfiler® includes AmpFISTR Y filer Primer sets (15-30bp) that are locus-specific, and AmpliTaq Gold® DNA Polymerase (an enzyme responsible for DNA replication which is obtained from expression of Thermos Aquatics DNA polymerase gene cloned in E.coli). The other components of the DNA amplification assay is AmpFISTR Y filer PCR Reaction Mix which contains magnesium chloride as a cofactor for polymerase, deoxyribonucleotide triphosphates (dNTPs), bovine serum albumin to stabilize the polymerase, sodium azide as preservative, TE buffer to maintain optimum buffering capacity, AmpFISTR Control DNA 007, and AmpFISTR Y filer allelic ladder.

2.6.3.2 Multiplex PCR Protocol:

All of the DNA amplification reactions were performed in multiplex fashion. The Multiplex PCR technology was introduced in 1988 (Chamberlain, 1988). Multiplexing PCR reaction allows simultaneous amplification of more than one locus in a single tube under the same conditions. In this study amplification of the 17 loci and amelogenin locus were performed according to manufacturer's protocol with

minor modification to adapt for automation. Multi-mix solution was prepared for a whole batch that contained the following volume as shown in table 2.

Table 2: Represent total volume for multi-mix of 96 reactions.

Plate Size	Reaction Mix (µl)	Primer (µl)	TaqGold (µl)
96	984 (ng/ µl)	515 (ng/ µl)	46 (ng/ µl)

Multi-Mix reaction tubes were vortexed for a minimum of 10 seconds and briefly spun to 5000rpm to ensure homogeneity of the components. The number of reactions per batch was 96 including positive and negative controls. Fifteen microliters of Multi-mix was added to each well, followed by the addition of 10 µl of DNA (1.5 ng/10µl) to a total volume of (25µl). The plate then was sealed by aluminum foil and centrifuged for 10 seconds and placed in the Bio-Rad Tetrad for high-throughput PCR applications.

2.6.3.3 Thermal Cycling Parameters for AmpFLSTR® Yfiler® PCR kit:

All PCR reactions consisted of hot-start cycle at 95° C for 11 minutes, followed by 29 cycles of denaturation at 94°C, annealing at 59° C for 1 minute, and extension at 72° C for 1 minute per cycle. The PCR reactions were completed by post-extension for 45 minutes at 60° C as shown in table 3.

Table 3: Demonstrate the Polymerase Chain Reaction cycling parameters.

Cycles	Stage	Temperature	Duration	Description
Once	Hot Start	95°C	11 min.	Reconfiguration of enzyme
28	Denaturation	94°C	1 min.	dsDNA strands split to ssDNA
	P. Annealing	59°C	1 min.	Primers anneal to binding site
	P. Extension	72°C	1 min.	Primers extended by enzyme
Once	Post- extension	60°C	45 min.	All strands extended to include extra Adenine base
Once	Hold	4°C	∞	Temperature decrease to denote end of process

2.6.3.4 Loci amplified by AmpFLSTR® Yfiler® kit:

The following table shows the loci amplified on Y chromosome, and the corresponding fluorescent marker dyes. The AmpFlSTR® Yfiler® Allelic Ladder is used to genotype and score the samples. The allele sizes represented in the allelic ladder and the genotype of the AmpFlSTR® Control DNA 007 are also listed in table 4.

Table 4: AmpFLSTR® Yfiler® kit loci and alleles (The table is adopted from manufacture protocol).

Locus designation	Alleles included in AmpFℓSTR® Yfiler® Allelic Ladder	Dye label	Control DNA 007
DYS456	13, 14, 15, 16, 17, 18	6-FAM™	15
DYS389 I	10, 11, 12, 13, 14, 15		13
DYS390	18, 19, 20, 21, 22, 23, 24, 25, 26, 27		24
DYS389 II	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34		29
DYS458	14, 15, 16, 17, 18, 19, 20	VIC®	17
DYS19	10, 11, 12, 13, 14, 15, 16, 17, 18, 19		15
DYS385 a/b	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25		11,14
DYS393	8, 9, 10, 11, 12, 13, 14, 15, 16	NED™	13
DYS391	7, 8, 9, 10, 11, 12, 13		11
DYS439	8, 9, 10, 11, 12, 13, 14, 15		12
DYS635	20, 21, 22, 23, 24, 25, 26		24
DYS392	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18		13
Y GATA H4	8, 9, 10, 11, 12, 13	PET®	13
DYS437	13, 14, 15, 16, 17		15
DYS438	8, 9, 10, 11, 12, 13		12
DYS448	17, 18, 19, 20, 21, 22, 23, 24		19

2.7 Detection of AmpFLSTR® Yfiler® kit PCR Product:

The multiplex amplified DNA fragments are resolved on a polymer-based electrophoresis technique in which the laser beam activates different fluorophase associated with each DNA fragments.

2.7.1 Principle of electrophoresis:

Electrophoretic separation is based on the principle that a charged particle in a solution will migrate towards one of the electrodes when placed in an electrical field. In our experiments, the DNA fragments were resolved by capillary electrophoreses (CE). The fundamental principle of CE is identical to the concept of gel electrophoreses. Briefly, the speed and direction a charged particle moves is determined by its net charge, size, shape and molecular weight as well as external factors such as the composition of the buffer, the voltage used among other variables. The fragments migrates across CE, where at a specific point along the capillary length the argon ion laser passes through clear sections in the array and excites the fluorescent dyes attached to the DNA fragments which result in emission of light at different detectable wavelengths. The specific wavelength of the emitted light is different for each dye. The light is collected and separated according to its wavelength by a spectrograph onto a charge coupled device (CCD) camera, and all types of fluorescent emissions can be detected with one burst of the laser. The data collection software collects the light intensities from specific areas on the CCD camera, corresponding to the different wavelengths of light. This is similar to using a physical filter to separate the light wavelengths, referred to as virtual filter.

2.7.2 Genetic Analyzer:

The separation and detection of AmpFLSTR® Yfiler® PCR kit was performed using the 3500XL Genetic Analyzer (Life Technologies). It is highly automated system and easy-to-use format allows for its wide usage in the laboratory. It Includes a 24 capillary 36cm in length, it analyses 24 samples every 25-30 minutes. In addition, it processes one sample per run as to reduce cross contamination. The fragments resolution are rather quick due to the high voltages. Previously this was not possible mainly due to heat generated at higher voltages which ultimately hampered the DNA fragment migration. Capillary has a large surface area (volume ratio), that allows efficient heat dissipation that reduces the time required for DNA fragment resolution. Moreover, there is 3500XL Data Collection Software, which is a Dell-based workstation and monitors as the parameter of the CE.

2.7.3 Capillary Electrophoresis Reagents:

1. Performance Optimized polymer (POP-4):

Capillary electrophoresis (CE) uses the same principles mentioned above. The samples are separated as they pass through a fine bore silica tube or (capillary) containing a liquid separation medium called Performance Optimized Polymer (POP-4). POP-4 reagent is a separation medium and it used by capillary electrophoresis POP-4 is the most suitable medium for DNA fragment resolution as it is non-cross linked liquid polymer. POP-4 contains denaturant, such as urea and Pyrrolidone that keeps DNA single stranded during resolution. The polymer is made of a non-cross linked / Acrylamide (N, N-dimethylacrylamide). Other reagents that are essential for capillary electrophoresis circuit are Anode (positive charge) and Cathode (negative

charge) buffer. Both Anode Buffer and part of the Cathode Buffer contain Tris-based EDTA and Boric acid buffer. It is required mainly for two reasons: first, it has a high pH 8.3, which allows DNA to exhibit a net negative charge. Second, it creates a complete electrical circuit allowing electrophoresis to take place.

2.7.4 Sample preparation before Capillary Electrophoresis:

Sample prepraration involves a couple of steps: the first is denaturation of double stranded amplified DNA to single strands and the second is the preparation of appropriate references for fragment scoring. They are:

1- Hi-Di-Formamide:

Hi-Di Formamide is a highly de-ionized form of the denaturing agent formamide. Formamide is a nucleic denaturant and changes, in this case, the DNA or RNA conformation to single stranded molecules DNA denaturation is an important component of analysis because single stranded DNA reduces the biases mobility shift due to DNA conformation and enhance analytical resolution.

2- GeneScanTM 600 LIZ ® Size Standard:

DNA fragment scoring is based on reference standard fragment. It is a solution of dye with labeled DNA fragments of known and varying lengths that is produced from bacterial plasmid restriction enzyme digestion. This results in a generation of fragments for size references that in conjunction with the allelic ladder allows verification of unknown STR lengths. In other words, it acts as an internal ruler for each sample allowing the size of the unknown fragments in the sample to be calculated. Figure 6 illustrates the fluorescent dye label color and relative PCR product size ranges for the various STR loci present in this particular kit.

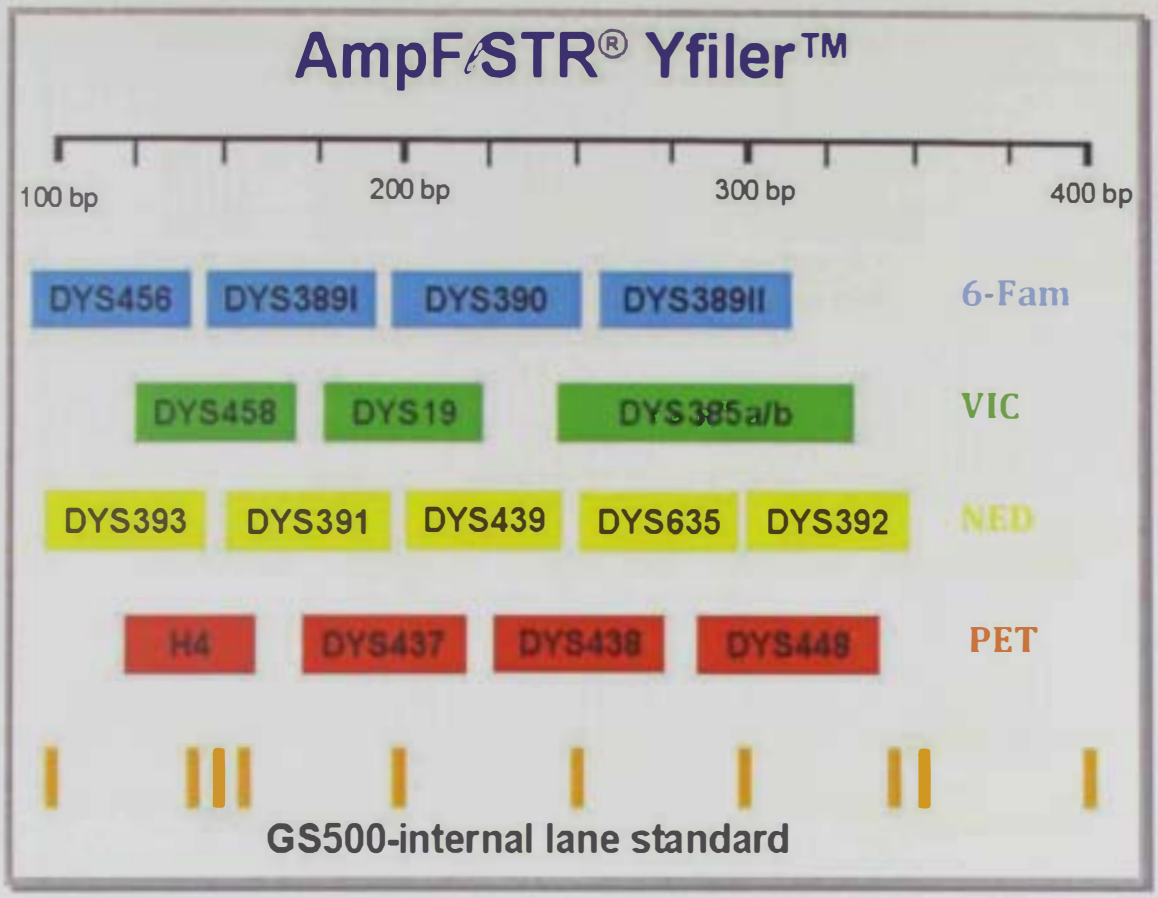


Figure 6: 17 loci of AmpFLSTR® Yfiler® PCR with ladder and internal control.

2.7.5 AmpFlSTR Y filer Allelic Ladder:

Fragment calling accuracy is reinforced by AmpFLSTR® Yfiler allelic ladder and internal control. The ladder consist of DNA fragment with known designation of the most commonly found alleles per locus as shown in Figure 7. These standard allele sizing are obtained from large population base study.

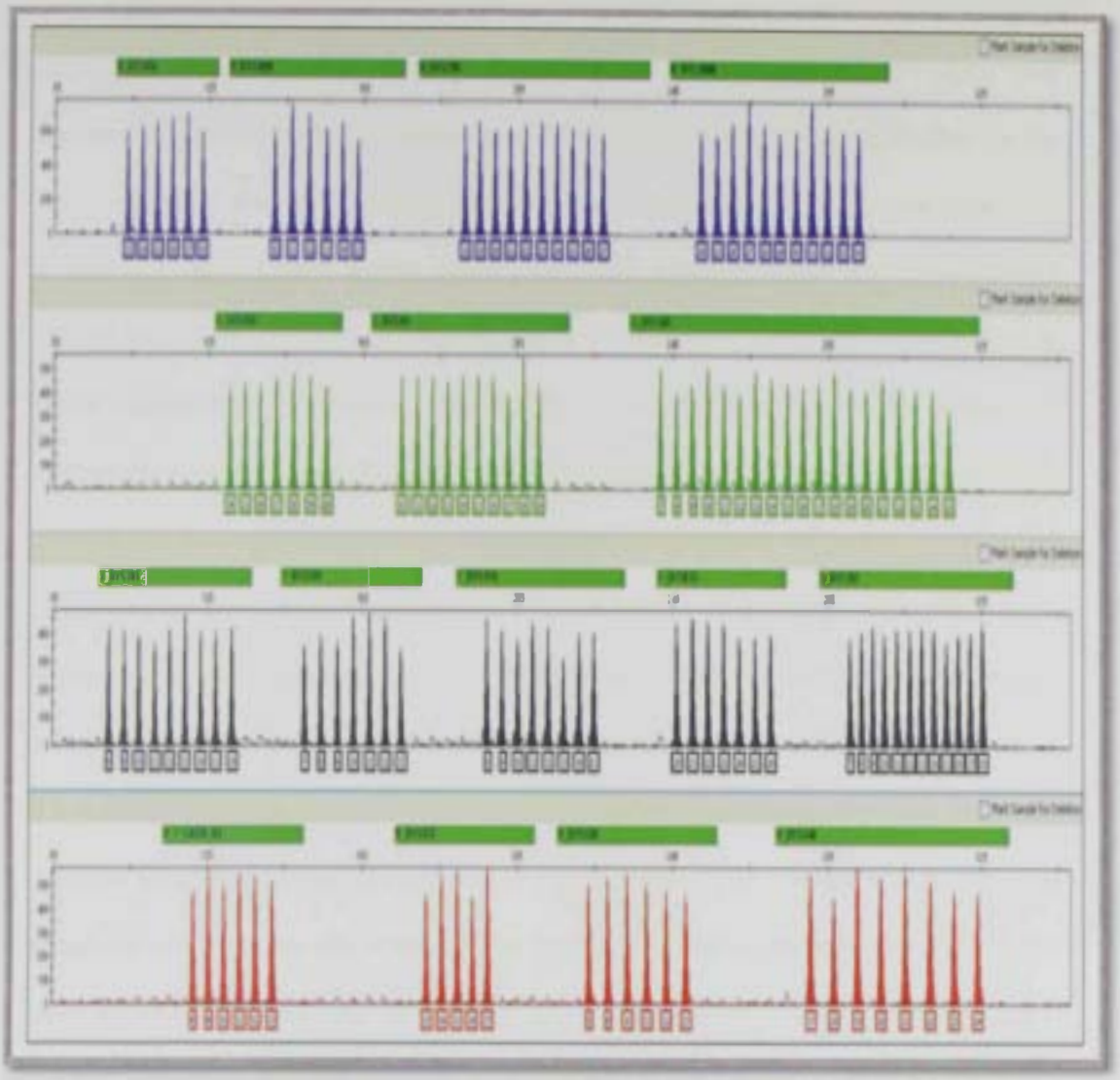


Figure 7: GeneMapper ID-X software plot the AmpFlSTR Yfiler Allelic Ladder.

Table 5: Total volumes required for DNA amplified plate in a CE step.

	Cocktail		
Plate Size	HiDi (µl)	Size Standard (µl)	Allelic Ladder(µl)
96	1404	36	1.5

HiDi and size standard reagents (cocktail) were mixed, vortexed, and spun in sterile microfuge tube for 10 seconds. Each sample consist of 13.5ul from the cocktail and 1.5ul from Allelic Ladder as shown in table 5. The plate was then pulse-spunned and

firmly seated with a cover slip septa®. Then the plate was subjected to heat (for denaturation of DNA) for 2 minutes at 95°C in thermo-cycler without closing the lid and then snaps cooled by placing the plate on a ice block. Finally, the plate was placed in a special plate holder and plate retainer placed on top, and the assembly placed onto the sequencer and started the run. The 3500 Data Collection software works alongside other software to control the mechanical operation of the instrument, such as moving the auto sampler and switching on the oven for DNA denaturation. It collects the fluorescence emission data from the CCD camera and processes it before storage as both tables (in the machines own database) and as sample files on the hard drive. Within the collection software, there are numerous modules containing pre-set instructions to the sequencer, governing parameters such as voltage, temperature of the oven, and the laser power. From the run window, the appropriate module can be selected for the plate check, pre-run and the run step. The collection software also monitors and displays the status of the instrument and saves it to the instrument database as EPT data.

2.8 Statistical Analysis:

For statistical analysis, UAE population was first divided into three regions, Northern that include; (Dubai, Sharjah, Ajman, Ras Alkhaima and Um Alqawin), Eastern which include (Fujurah, Kalbaa and Khorfakkan) and Western region include (Abu Dhabi and AlAin city) and interchangeably they are called population 1, 2, and 3 respectively. The population was grouped based on the history, geographic proximities, and probable migration routes. The analysis of allele frequency and haplotype were performed with statistical and population genetic software.

GeneMapper® ID-X version 1.2 analysis software carried out data analysis. It analyzes data generated on human identification capillary electrophoresis platforms such as 3500XL Genetic Analyzer. In this project, the usage of this software was to analyze the results and to generate raw data for all samples and the parameter was set to 150 Relative Fluorescent Unit (rfu). The raw data file that generated from the GeneMapper® ID-X was analyzed using macro programming in Microsoft Excel and an integrated software package for population genetics data analysis ALEQUIN version 3.5.1.3. (Excoffier, 2005). Moreover, statistical program SPSS (Statistical Package for Social Sciences) was used to generate the descriptive parameter of population such as Mean, Standard deviation, allele frequency, charts and tables for each loci.

2.9 Analysis of Population genetic parameters:

The field of population genetics has come a long way since the early part of the 20th century. The combination of Mendelian genetics and biometric studies led to the birth of population genetics whose father is R. Fisher (Hald, 1998). Initially, population genetics studies concentrated on studying the allele frequencies in populations, subpopulation, and groups (Provine, 1978). Therefore, it is important to develop the notion of allele frequencies computation even though our current study involves a haploid system manifested in Y chromosomal STR analysis. It is important to describe Hardy and Weinberg Equilibrium (HWE) as many of the mathematical treatments in population genetics stems from HWE.

The HWE method is used to calculate the expected proportion of different genotypes in a given population. A population to which the HWE is applicable

consists of diploid systems, sexually reproducing individuals, and a score of other assumptions such as independence of allele frequencies transmission from one generation to the next from the common genetic pool who has not been under any selection force. In the HWE, the probabilities of the genotypes can be given by the equation:

$p^2 + 2pq + q^2 = 1$,

Where: p^2 = percentage of homozygous dominant individuals

q^2 = percentage of homozygous recessive individuals

2pq = percentage of heterozygous individuals;

And $p + q = 1$

Where:

p = frequency of the dominant allele in the population and

q = frequency of the recessive allele in the population.

The total frequency of alleles in a population is equal to one. In addition, HWE depends mainly on the existence of a very large size and randomly mating populations (Hammond et al., 1994).

There are two important facets to Hardy-Weinberg model. First and the most important facet, is that it shows that the Mendelian mechanism preserves genetic variability. Second, it provides a useful functional relationship between genotype frequencies and gene (allele) frequencies where $p^2 + 2pq + q^2 = (p+q)^2 = 1$. It shows that, everything else being equal, the population will quickly reach equilibrium and stay there. However, HWE works in an ideal situation, that does not usually occur naturally and thus HWE is used as a null model to test various population structure assumptions such genetic forces and population constraints.

The second point which is mostly important to population geneticists is the ability to describe the state of a population entirely in terms of allele frequencies rather than genotype frequencies. Allele (which are much fewer than genotypes) frequencies of the population can derive the expected heterogeneity of population.

2.10 Measure of diversity between and within population:

Over the past few decades, microsatellites have been the choice for population studies, due to their high level of variability and the relative ease of development and scoring in non-model systems. Several methods have been used to score the variations and diversity of microsatellite in populations. *Fst* or fixation index is one of the most used biometric method. *Fst* simply measures the level of hetrozygosity or differentiation between population and its subpopulation. It was originally developed to measure genetic distance using biallelic markers (Wright, 1969), but the equation was subsequently generalized for multiple alleles (Nei, 1973). F_{St} ranges from 0.0 to 1.0, as expected, with 0 indicating no differences in allele frequencies between two populations and 1.0 indicating that the two populations are fixed for alternate alleles. F_{St} is often 0.1-0.2. For microsatellites with high mutation rate (in the range of 10^{-2}-10^{-6}) other measures of hetrozygosity or diversity has been introduced to study population structure. Subsequent mathematical derivation, namely, R_{ST}, is used to account for mutation under the assumption of step-wise mutations, that is, single mutation at a time and that each subsequent mutation is dependent on the previous one-that is the size of microsatellite dictates the future mutation (Slatkin, 1997). The previous population modeling eventually diverted from mere allelic frequency studies to genetic distance among and between populations which is mainly derived from

allelic variance between subpopulations. In the current study, we have estimated haplotype frequencies, variances, and used analysis of molecular variance (AMOVA) to study the UAE population and its subgroups. Furthermore, we have compared the UAE population with other populations in different parts of the world. Our analysis was conducted on ALEQUIN (version 3.5.1.3.) which is an integrated software package for population genetics data analysis (Excoffier, 2005) and SPSS statistical package.

CHAPTER III: RESULTS AND DISCUSSION

This chapter describes the results of the DNA profiling of 345 samples obtained from unrelated males in the UAE. All of the samples were genotyped with AmpFl STR® Y filer™ kit (Life Technology) for 17 highly polymorphic loci. This is one of the largest single population studied thus far.

3.1 Profiling Samples:

All the 345 samples were successfully profiled for all 17 loci with the AmpFl STR® Y filer™ kit. Figure 8 presents an example of Y STR profile for a subject. The same reaction was performed for 345 samples.

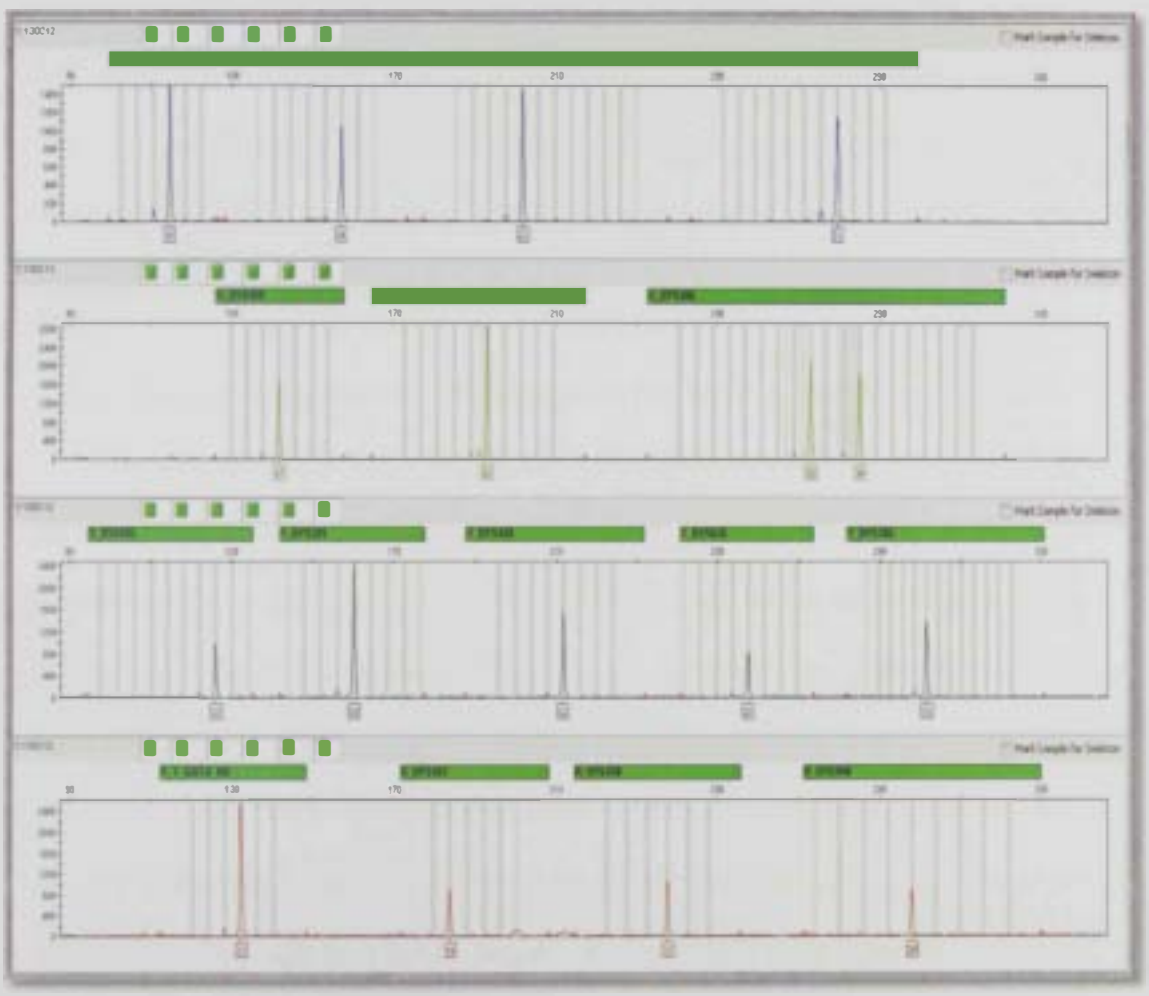


Figure 8: Y STR profile generated using AmpFl STR® Y filer™ at optimal reaction conditions. The figure shows labelled individual STR and the peak height.

3.2 Y STR alleles Frequencies:

In general, allele frequencies are used to study the population structure and its variance is a key indicator of genetic diversity at the individual, population, and species level and demonstrates the richness of the population gene pool (Heuertz et al., 2004). The population genetic parameters are computed with statistical package for the social science (SPSS) and population genetics software Arlequin (Excoffier, 2005) for all 17 amplifiable loci. The computations are carried out for the whole population as well as each subpopulation (sections 3.2.1 and 3.2.2 respectively). It is important to note that in this study, the state of hetrozygosity of allele frequencies of 345 Emirati samples were estimated by counting the number of individual allele observed and dividing it by the total number of samples for the respective population. We have observed a total number of 124 alleles distributed across the 17 loci in the UAE population (see table 6).

3.2.1 Allele Frequency of Y Haplotype in the UAE population:

The analysis of the allele frequency in the UAE population clearly shows that each locus has a predominant allele (see table 7 and figure 9). It is also apparent that the alleles of most loci are clustered over a narrow range where approximately 60% - 80% of the population is sharing a specific allele for the locus. For example, allele 14 of DYS437 is shared among 79% of the population; similarly allele 11 of DYS392 is shared among 79% of the UAE population, allele 13 of DYS3891 locus is shared in 72% of the population, and allele 10 of DYS391 is shared between 72% of the population. Although to a lesser degree, other loci also have predominant alleles.

These loci are DYS438 (allele 10 is shared among 59% of population), DYS448 (allele 20 is shared among 56% of population), DYS19 (allele 14 is shared among 55% of population), GATA_H4 (allele 11 is shared among 54% of population), DYS393 (allele 11 is shared in 48% of population), and DYS456 (allele 15 is shared in 47% of population).

In other cases, there is a bimodal or even multimodal distribution of alleles (see table 8; i.e. DYS458 locus); that is alleles that are not clustered and their distribution appears discontinuous and chunky. For instance, DYS458 locus shows two predominant alleles (16 and 17) with frequencies of 68 and 72 respectively (where N= 345). While the most common alleles at DYS439 locus are 11 and 12 with frequencies of 151 and 118 respectively. Moreover, for DYS458 locus the two common alleles are 17 with the frequency of 72 and allele 16 with the frequency of 68 and for the DYS385-B locus the predominant alleles are 18 and 17 with frequencies of 88 and 68 respectively. The bimodal allele predominance distribution is characteristic of gene flow from other populations. That is there are groups of the population, arguably, that migrated into the UAE from different regions and further studies are required to identify the most likely alleles that are associated with each allelic modal. Table 8 (A-Q) presents a detailed description of allelic distribution for each loci and the corresponding pie charts.

Table 6: Total number of alleles for each loci in UAE population

Locus #	# of alleles
DYS456	5
DYS389I	4
DYS390	6
DYS389II	8
DYS458	13
DYS19	6
DYS385-A	11
DYS385-B	11
DYS393	8
DYS391	7
DYS439	7
DYS635	9
DYS392	7
GATA_H4	5
DYS437	4
DYS438	5
DYS448	8
Mean	7.294
S.D.	2.568

Table 7: Predominant allele in UAE population.

Locus	Predominant	Frequency
DYS456	15	0.475
DYS3891	13	0.722
DYS390	23	0.441
DYS389II	30	0.432
DYS458	17.16*	0.209, 0.197*
DYS19	14	0.554
DYS385-A	13	0.377
DYS385-B	18	0.255
DYS393	12	0.475
DYS391	10	0.643
DYS439	11	0.438
DYS635	21	0.417
DYS392	11	0.783
GATA_H4	11	0.539
DYS437	14	0.786
DYS438	10	0.586
DYS448	20	0.565

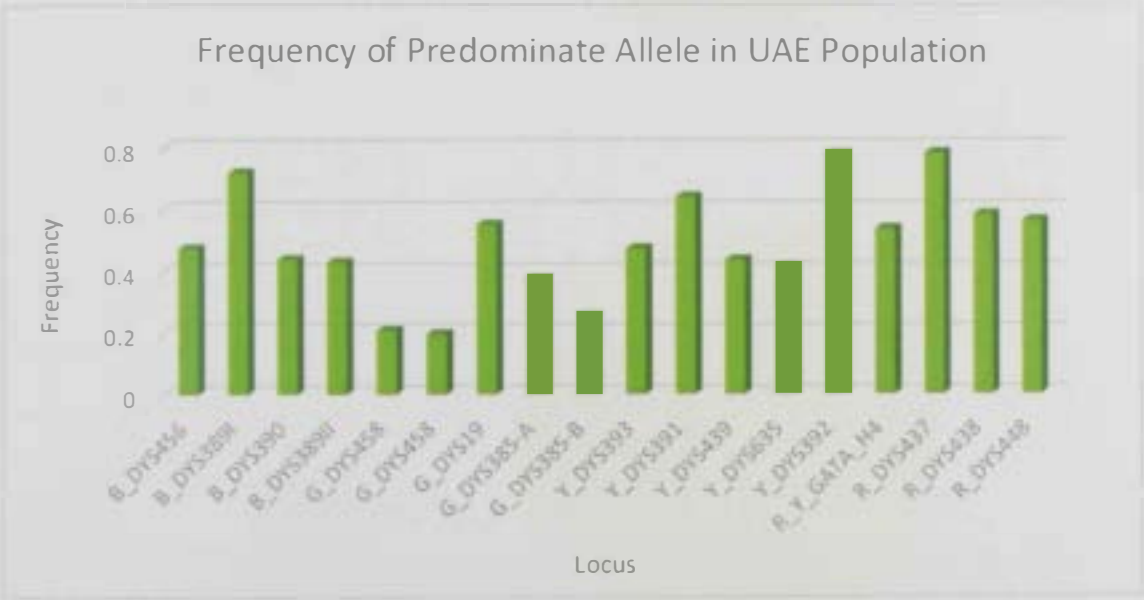
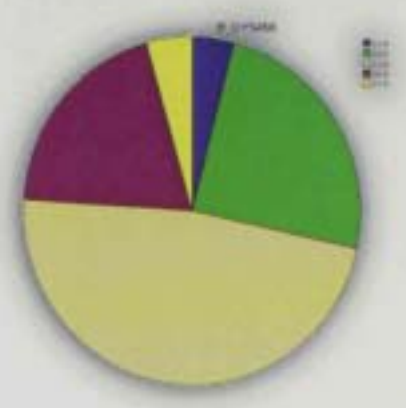


Figure 9: Allele frequency for the predominant alleles in the UAE population

Table 8: Allele frequency of different loci with pie charts for UAE population:

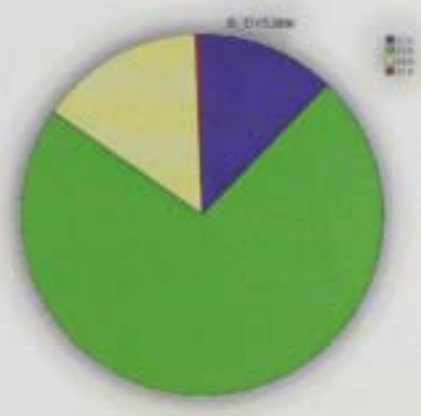
A. Locus DYS456

Allele	Frequency	Percent
13.0	14	4.1
14.0	84	24.3
15.0	164	47.5
16.0	68	19.7
17.0	15	4.3
Total	345	100.0



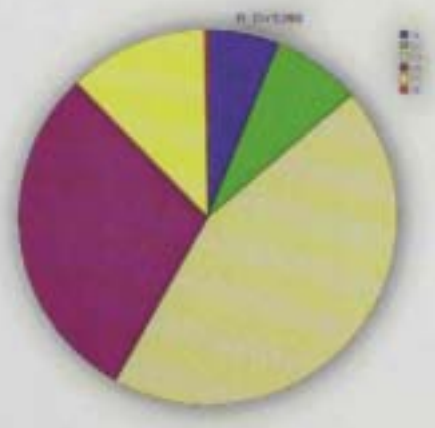
B. Locus DYS3891

Allele	Frequency	Percent
12.0	42	12.2
13.0	249	72.2
14.0	51	14.8
15.0	2	.6
Total	345	100.0



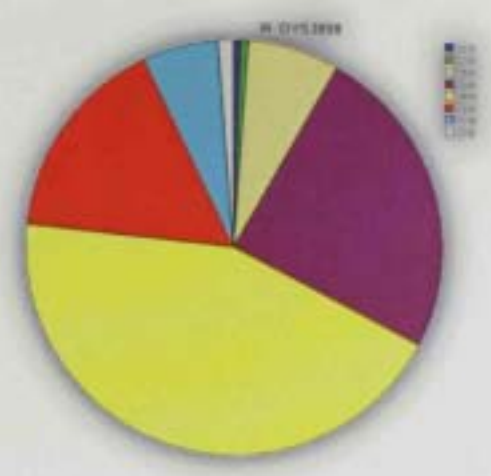
C. Locus DYS390

Allele	Frequency	Percent
21	21	6.1
22	26	7.5
23	152	44.1
24	101	29.3
25	41	11.9
26	1	.3
Total	345	100.0



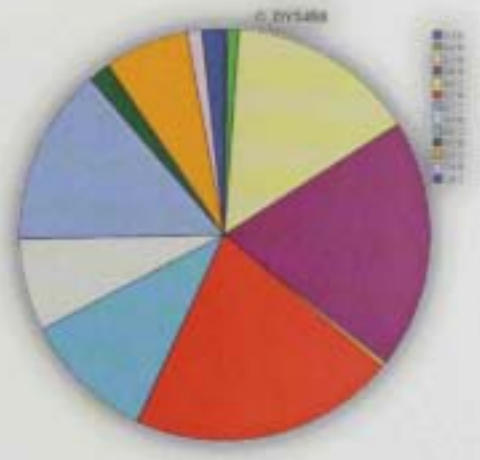
D. Locus DYS38911

Allele	Frequency	Percent
25.0	2	.6
27.0	2	.6
28.0	24	7.0
29.0	83	24.1
30.0	149	43.2
31.0	55	15.9
32.0	20	5.8
33.0	4	1.2
Total	345	100.0



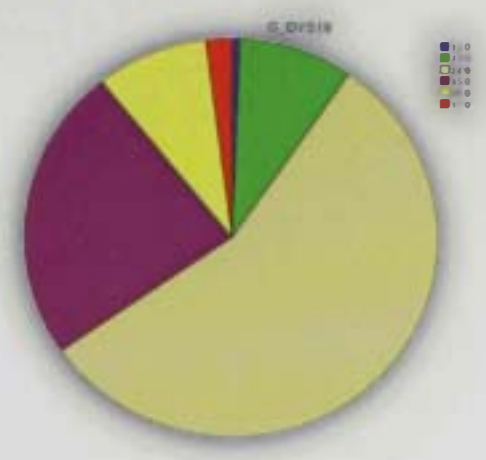
E. Locus DYS458

Allele	Frequency	Percent
13.0	1	.3
14.0	3	.9
15.0	51	14.8
16.0	68	19.7
16.2	1	.3
17.0	72	20.9
17.2	36	10.4
18.0	25	7.2
18.2	48	13.9
19.0	6	1.7
19.2	23	6.7
20.0	4	1.2
20.2	6	1.7
Total	345	100.0



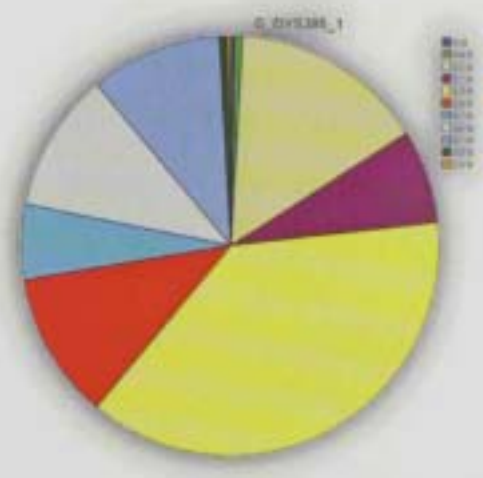
F. Locus DYS19

Allele	Frequency	Percent
12.0	2	.6
13.0	31	9.0
14.0	191	55.4
15.0	82	23.8
16.0	30	8.7
17.0	7	2.0
Total	345	100.0



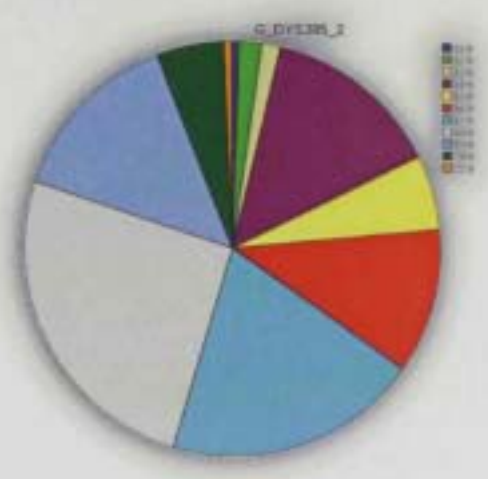
G. Locus DYS385_1

Allele	Frequency	Percent
8.0	1	.3
10.0	2	.6
11.0	52	15.1
12.0	25	7.2
13.0	130	37.7
14.0	39	11.3
15.0	20	5.8
16.0	37	10.7
17.0	35	10.1
18.0	2	.6
19.0	1	3
Total	345	100.0



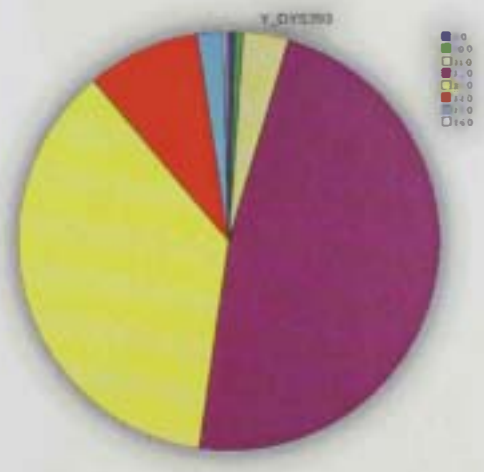
H. Locus DYS385_2

Allele	Frequency	Percent
11.0	2	.6
12.0	6	1.7
13.0	5	1.4
14.0	48	13.9
15.0	20	5.8
16.0	39	11.3
17.0	68	19.7
18.0	88	25.5
19.0	48	13.9
20.0	18	5.2
21.0	2	.6
Total	345	100.0



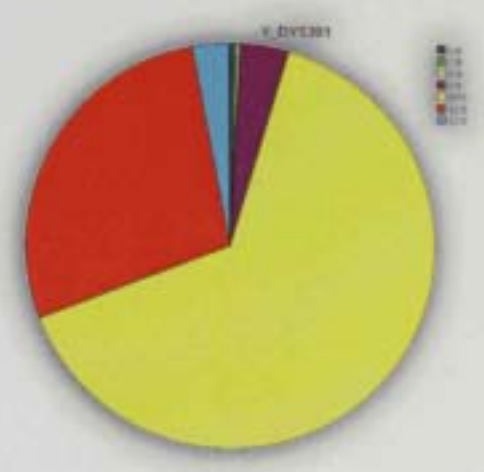
I. Locus DYS393

Allele	Frequency	Percent
9.0	2	.6
10.0	2	.6
11.0	12	3.5
12.0	164	47.5
13.0	126	36.5
14.0	30	8.7
15.0	7	2.0
16.0	1	.3
Total	345	100.0



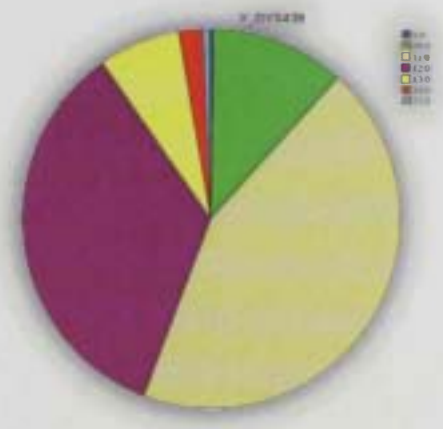
J. Locus DYS391

Allele	Frequency	Percent
6.0	1	.3
7.0	1	.3
8.0	1	.3
9.0	13	3.8
10.0	222	64.3
11.0	96	27.8
12.0	10	2.9
Total	345	100.0



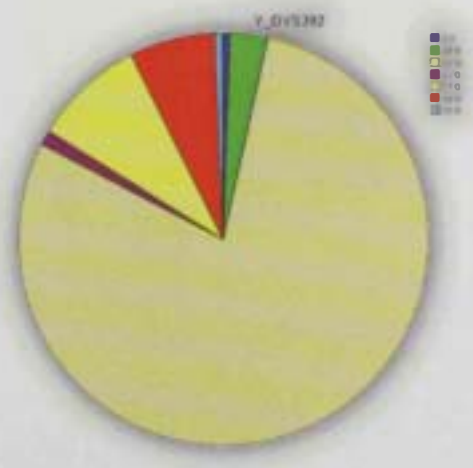
K. Locus DYS439

Allele	Frequency	Percent
9.0	1	.3
10.0	39	11.3
11.0	151	43.8
12.0	118	34.2
13.0	24	7.0
14.0	7	2.0
15.0	2	.6
Total	345	100.0



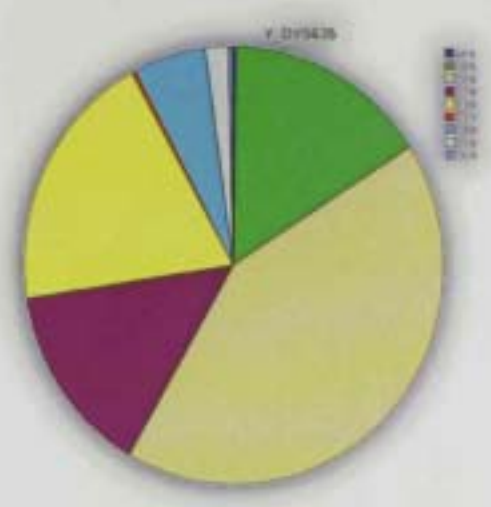
L. Locus DYS392

Allele	Frequency	Percent
9.0	2	.6
10.0	10	2.9
11.0	270	78.3
12.0	4	1.2
13.0	30	8.7
14.0	23	6.7
15.0	2	.6
Total	345	100.0



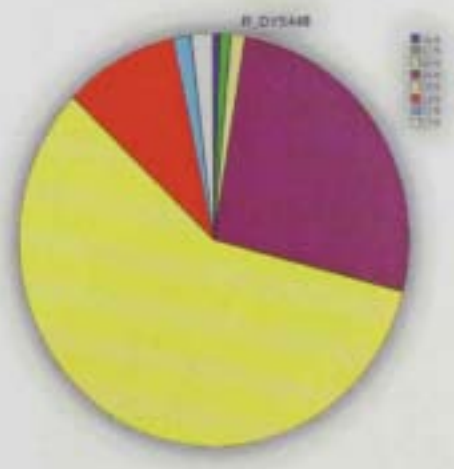
M. Locus DYS635

Allele	Frequency	Percent
19.0	1	.3
20.0	53	15.4
21.0	144	41.7
22.0	48	13.9
23.0	66	19.1
23.3	1	.3
24.0	19	5.5
25.0	6	1.7
26.0	1	.3
Total	345	100.0



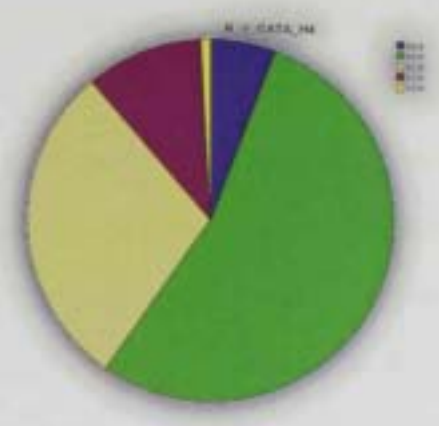
N. Locus DYS448

Allele	Frequency	Percent
16.0	2	6
17.0	3	9
18.0	4	1.2
19.0	89	25.8
20.0	195	56.5
21.0	32	9.3
22.0	5	1.4
23.0	6	1.7
Total	345	100.0



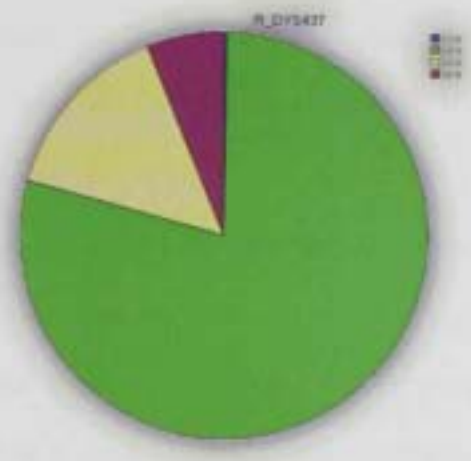
O. Locus Y_GATA_H4

Allele	Frequency	Percent
10.0	20	5.8
11.0	186	53.9
12.0	100	29.0
13.0	36	10.4
14.0	3	.9
Total	345	100.0



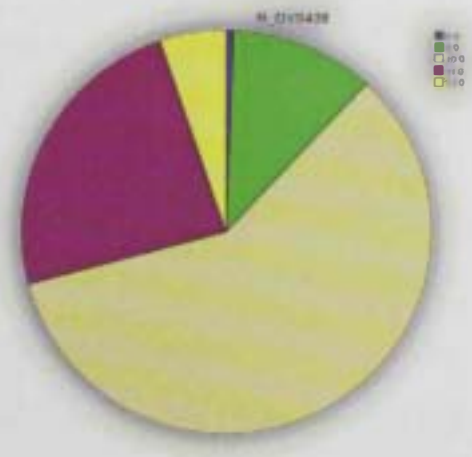
P. Locus DYS437

Allele.	Frequency	Percent
13.0	1	.3
14.0	271	78.6
15.0	50	14.5
16.0	21	6.1
Total	345	100.0



Q. Locus DYS438

Allele	Frequency	Percent
8.0	2	.6
9.0	39	11.3
10.0	202	58.6
11.0	82	23.8
12.0	18	5.2
Total	345	100.0



It is evident from the predominant allele frequencies and bimodality distribution of several alleles at specific locus that there are at least minimal common haplotypes comprising UAE population. In fact, our analysis shows that approximately half of the population share at least 5-7 alleles across the 17 STR loci (see section 3.2.2 for further discussion). Based on this evidence, two scenarios are viable to explain the structure of the UAE population. First, a steady population increase in the past several thousand years within the boundaries of what constitutes the UAE. The initial population shared common haplotype that eventually mutated to establish the genetic diversity of today's population. The second scenario involves recent admixture and gene flow for multiple neighboring regions within the past century. Although the second scenario is the most plausible, it is important to stress the importance of covering a larger population at higher population hierarchy within the region at higher molecular resolution to decipher the population structure and establish deep ancestry (Underhill, 2007). Alternatively, both scenarios at work although the first scenario provided the diversity of alleles, contributed to a lesser degree to the haplotypes diversity. This latter assertion stems from an observation of a narrow range of alleles clustering within each locus (see table 8). The allele clustering and narrow range distribution assumes a stepwise mutation as opposed to infinite allele model (Valdes et.al. 1993; Jarne, et. al. 1996).

Moreover, the UAE population, as is the case in any other population, demonstrates allele frequencies and genetic variance that is distinct from other populations in different continents. The result reflects the genetic distance between these populations, which implies the separation of population inhabiting the UAE

regions and the surrounding area at least in the past several thousand years (Eckert and Hile 2009; Cadenas, 2008; Bosch, 2000). Diversity of this magnitude in a relatively short period (several thousand years) is in agreement with higher mutation rate of microsatellite (10^{-2}-10^{-4} per generation) as opposed to single nucleotide mutation (Brinkmann et al., 1998; Dupuy et al., 2004).

3.2.2 Allele Frequency for subpopulations in the UAE:

The allelic profile, frequency, and distribution for each locus in the three subpopulations of the UAE follows similar patterns (see table 9 and figure 10.A-Q). The allele frequency among the three subpopulations, however, shows small fluctuations especially for specific loci. Although, our study demonstrates that there are unique alleles to a subpopulation or alleles not shared by all regions, the profile of the allele frequency and distribution are similar. It is important to note that the sample size of Eastern subpopulation (n=25) is smaller compared to Western (n=189) and Northern (n=139) regions; therefore, we attempt to discuss the results of Eastern region with caution even though it is not uncommon to see a population of similar size discussed for inferences in the literature. Table 10 demonstrates the alleles that are unique or not shared between all the regions in the UAE population. For example, allele 15 of DYS389I, is present only in Western region, while alleles 26 of DYS390 is present in Northern region only. Interestingly, alleles 13 and 16.2 of DYS458 locus and alleles 9 and 16 of DYS393 locus are present only in Western region. Moreover, alleles (10 and 19) of DYS385-A locus are unique to Northern region only while alleles 8 and 18 are only present in Western region. We have excluded the Eastern region in this comparison as the population size is smaller

(n=25) than Northern and Western regions. However, the absence of an allele from a region with high population does warrant its absence in the Eastern region population (Table 10) that there are alleles only specific to Eastern region. A larger population size will ultimately identify the uniqueness of these alleles in the respective subpopulations.

Table 9: Allele Frequencies for the sub-populations.

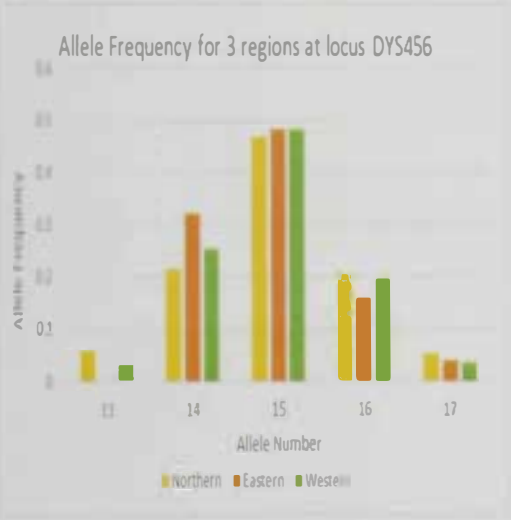
Northern (n=131)				Eastern (n=25)				Western (n=189)			
Alleles for the locus 1: DYS456:											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.061	0.021	13.0	1	0.000	0.000	13.0	1	0.032	0.007	13.0
2	0.214	0.036	14.0	2	0.320	0.067	14.0	2	0.254	0.018	14.0
3	0.466	0.044	15.0	3	0.480	0.071	15.0	3	0.481	0.021	15.0
4	0.206	0.035	16.0	4	0.160	0.052	16.0	4	0.196	0.017	16.0
5	0.053	0.020	17.0	5	0.040	0.028	17.0	5	0.037	0.008	17.0
Alleles for the locus 2: DYS3891:											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.122	0.029	12.0	1	0.040	0.028	12.0	1	0.132	0.014	12.0
2	0.763	0.037	13.0	2	0.680	0.067	13.0	2	0.698	0.019	13.0
3	0.107	0.027	14.0	3	0.280	0.064	14.0	3	0.159	0.015	14.0
4	0.000	0.000	15.0	4	0.000	0.000	15.0	4	0.011	0.004	15.0
Alleles for the locus 3: DYS3890:											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.061	0.021	21.0	1	0.080	0.039	21.0	1	0.058	0.010	21.0
2	0.053	0.020	22.0	2	0.040	0.028	22.0	2	0.095	0.012	22.0
3	0.420	0.043	23.0	3	0.520	0.071	23.0	3	0.444	0.021	23.0
4	0.282	0.039	24.0	4	0.280	0.064	24.0	4	0.302	0.019	24.0
5	0.168	0.033	25.0	5	0.040	0.028	25.0	5	0.095	0.012	25.0
6	0.008	0.008	26.0	6	0.000	0.000	26.0	6	0.000	0.000	26.0
Alleles for the locus 4: DYS389II											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.015	0.011	25.0	1	0.000	0.000	25.0	1	0.000	0.000	25.0
2	0.000	0.000	27.0	2	0.000	0.000	27.0	2	0.011	0.004	27.0
3	0.084	0.024	28.0	3	0.000	0.000	28.0	3	0.069	0.011	28.0
4	0.252	0.038	29.0	4	0.280	0.064	29.0	4	0.228	0.018	29.0
5	0.405	0.043	30.0	5	0.400	0.070	30.0	5	0.455	0.021	30.0
6	0.153	0.032	31.0	6	0.240	0.061	31.0	6	0.153	0.015	31.0
7	0.069	0.022	32.0	7	0.000	0.000	32.0	7	0.058	0.010	32.0
8	0.008	0.008	33.0	8	0.040	0.028	33.0	8	0.011	0.004	33.0
Alleles for the locus 5: DYS458											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	13.0	1	0.000	0.000	13.0	1	0.005	0.003	13.0
2	0.008	0.008	14.0	2	0.000	0.000	14.0	2	0.011	0.004	14.0
3	0.176	0.033	15.0	3	0.120	0.046	15.0	3	0.132	0.014	15.0

4	0.229	0.037	16.0	4	0.160	0.052	16.0	4	0.180	0.016	16.0
5	0.000	0.000	16.2	5	0.000	0.000	16.2	5	0.005	0.003	16.2
6	0.229	0.037	17.0	6	0.160	0.052	17.0	6	0.201	0.017	17.0
7	0.092	0.025	17.2	7	0.080	0.039	17.2	7	0.116	0.013	17.2
8	0.069	0.022	18.0	8	0.040	0.028	18.0	8	0.079	0.011	18.0
9	0.099	0.026	18.2	9	0.032	0.067	18.2	9	0.143	0.015	18.2
10	0.008	0.008	19.0	10	0.040	0.028	19.0	10	0.021	0.006	19.0
11	0.069	0.022	19.2	11	0.040	0.028	19.2	11	0.069	0.011	19.2
12	0.008	0.008	20.0	12	0.040	0.028	20.0	12	0.011	0.004	20.0
13	0.015	0.011	20.2	13	0.000	0.000	20.2	13	0.021	0.006	20.2
Alleles for the locus 6: DYS19											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	12.0	1	0.000	0.000	12.0	1	0.011	0.004	12.0
2	0.107	0.027	13.0	2	0.120	0.046	13.0	2	0.074	0.011	13.0
3	0.542	0.044	14.0	3	0.560	0.071	14.0	3	0.561	0.021	14.0
4	0.200	0.060	15.0	4	0.280	0.064	15.0	4	0.228	0.018	15.0
5	0.084	0.024	16.0	5	0.040	0.028	16.0	5	0.095	0.012	16.0
6	0.015	0.011	17.0	6	0.000	0.000	17.0	6	0.026	0.007	17.0
Alleles for the locus: 7 DYS385-A											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	8.0	1	0.000	0.000	8.0	1	0.005	0.003	8.0
2	0.015	0.011	10.0	2	0.000	0.000	10.0	2	0.000	0.000	10.0
3	0.198	0.035	11.0	3	0.120	0.046	11.0	3	0.122	0.014	11.0
4	0.076	0.023	12.0	4	0.200	0.057	12.0	4	0.053	0.009	12.0
5	0.351	0.042	13.0	5	0.360	0.069	13.0	5	0.397	0.021	13.0
6	0.084	0.024	14.0	6	0.080	0.039	14.0	6	0.138	0.014	14.0
7	0.038	0.017	15.0	7	0.120	0.046	15.0	7	0.063	0.010	15.0
8	0.099	0.026	16.0	8	0.120	0.046	16.0	8	0.111	0.013	16.0
9	0.122	0.029	17.0	9	0.000	0.000	17.0	9	0.101	0.013	17.0
10	0.000	0.000	18.0	10	0.000	0.000	18.0	10	0.011	0.004	18.0
11	0.008	0.008	19.0	11	0.000	0.000	19.0	11	0.000	0.000	19.0
Alleles for the locus 8: DYS385-B											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.008	0.008	11.0	1	0.000	0.000	11.0	1	0.005	0.003	11.0
2	0.023	0.013	12.0	2	0.000	0.000	12.0	2	0.016	0.005	12.0
3	0.008	0.008	13.0	3	0.000	0.000	13.0	3	0.021	0.006	13.0
4	0.160	0.032	14.0	4	0.200	0.057	14.0	4	0.116	0.013	14.0
5	0.084	0.024	15.0	5	0.000	0.000	15.0	5	0.048	0.009	15.0
6	0.099	0.026	16.0	6	0.080	0.039	16.0	6	0.127	0.014	16.0
7	0.183	0.034	17.0	7	0.160	0.052	17.0	7	0.212	0.017	17.0
8	0.267	0.039	18.0	8	0.280	0.064	18.0	8	0.243	0.018	18.0

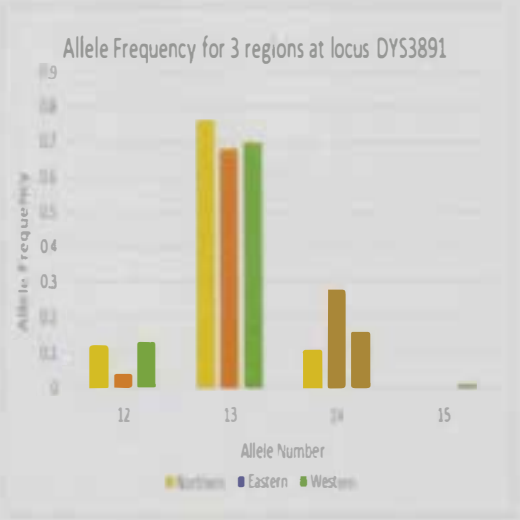
9	0.130	0.029	19.0	9	0.200	0.057	19.0	9	0.138	0.014	19.0
10	0.031	0.015	20.0	10	0.080	0.039	20.0	10	0.063	0.010	20.0
11	0.000	0.000	21.0	11	0.000	0.000	21.0	11	0.011	0.004	21.0
Alleles for the locus 9: DYS393											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	9.0	1	0.000	0.000	9.0	1	0.011	0.004	9.0
2	0.008	0.008	10.0	2	0.000	0.000	10.0	2	0.005	0.003	10.0
3	0.038	0.017	11.0	3	0.040	0.028	11.0	3	0.032	0.007	11.0
4	0.435	0.043	12.0	4	0.520	0.071	12.0	4	0.497	0.021	12.0
5	0.405	0.043	13.0	5	0.280	0.064	13.0	5	0.349	0.020	13.0
6	0.084	0.024	14.0	6	0.120	0.046	14.0	6	0.085	0.012	14.0
7	0.023	0.013	15.0	7	0.040	0.028	15.0	7	0.016	0.005	15.0
8	0.000	0.000	16.0	8	0.000	0.000	16.0	8	0.005	0.003	16.0
Alleles for the locus 10: DYS391											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	6.0	1	0.000	0.000	6.0	1	0.005	0.003	6.0
2	0.000	0.000	7.0	2	0.000	0.000	7.0	2	0.005	0.003	7.0
3	0.000	0.000	8.0	3	0.000	0.000	8.0	3	0.005	0.003	8.0
4	0.031	0.015	9.0	4	0.040	0.028	9.0	4	0.042	0.008	9.0
5	0.679	0.041	10.0	5	0.640	0.069	10.0	5	0.619	0.020	10.0
6	0.260	0.038	11.0	6	0.280	0.064	11.0	6	0.291	0.019	11.0
7	0.023	0.013	12.0	7	0.040	0.028	12.0	7	0.032	0.007	12.0
Alleles for the locus 11: DYS439											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	9.0	1	0.000	0.000	9.0	1	0.005	0.003	9.0
2	0.122	0.029	10.0	2	0.080	0.039	10.0	2	0.111	0.013	10.0
3	0.420	0.043	11.0	3	0.520	0.071	11.0	3	0.439	0.021	11.0
4	0.336	0.041	12.0	4	0.240	0.061	12.0	4	0.360	0.020	12.0
5	0.092	0.025	13.0	5	0.080	0.039	13.0	5	0.053	0.009	13.0
6	0.023	0.013	14.0	6	0.040	0.028	14.0	6	0.016	0.005	14.0
7	0.000	0.000	15.0	7	0.000	0.000	15.0	7	0.011	0.004	15.0
Alleles for the locus 12: DYS635											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.008	0.008	19.0	1	0.000	0.000	19.0	1	0.000	0.000	19.0
2	0.176	0.033	20.0	2	0.080	0.039	20.0	2	0.148	0.015	20.0
3	0.374	0.042	21.0	3	0.440	0.071	21.0	3	0.444	0.021	21.0
4	0.137	0.030	22.0	4	0.120	0.046	22.0	4	0.143	0.015	22.0
5	0.214	0.036	23.0	5	0.240	0.061	23.0	5	0.169	0.016	23.0
6	0.000	0.000	24.0	6	0.040	0.028	24.0	6	0.058	0.010	24.0
7	0.023	0.013	25.0	7	0.000	0.000	25.0	7	0.016	0.005	25.0

8	0.000	0.000	26.0	8	0.040	0.028	26.0	8	0.000	0.000	26.0
Alleles for the locus 13: DYS392											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	9.0	1	0.000	0.000	9.0	1	0.011	0.004	9.0
2	0.023	0.013	10.0	2	0.080	0.039	10.0	2	0.026	0.007	10.0
3	0.779	0.036	11.0	3	0.720	0.064	11.0	3	0.794	0.017	11.0
4	0.008	0.008	12.0	4	0.040	0.028	12.0	4	0.011	0.004	12.0
5	0.099	0.026	13.0	5	0.080	0.039	13.0	5	0.079	0.011	13.0
6	0.084	0.024	14.0	6	0.000	0.000	14.0	6	0.063	0.010	14.0
7	0.000	0.000	15.0	7	0.040	0.028	15.0	7	0.005	0.003	15.0
Alleles for the locus 14: GHTA-H4											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.092	0.025	10.0	1	0.040	0.028	10.0	1	0.037	0.008	10.0
2	0.481	0.044	11.0	2	0.680	0.067	11.0	2	0.561	0.021	11.0
3	0.290	0.040	12.0	3	0.080	0.039	12.0	3	0.317	0.020	12.0
4	0.137	0.030	13.0	4	0.120	0.046	13.0	4	0.079	0.011	13.0
5	0.000	0.000	14.0	5	0.080	0.039	14.0	5	0.005	0.003	14.0
Alleles for the locus 15: DYS437											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	13.0	1	0.000	0.000	13.0	1	0.005	0.003	13.0
2	0.763	0.037	14.0	2	0.800	0.057	14.0	2	0.799	0.017	14.0
3	0.168	0.033	15.0	3	0.080	0.039	15.0	3	0.138	0.014	15.0
4	0.061	0.021	16.0	4	0.080	0.039	16.0	4	0.058	0.010	16.0
Alleles for the locus 16: DYS438											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.000	0.000	8.0	1	0.040	0.028	8.0	1	0.005	0.003	8.0
2	0.107	0.027	9.0	2	0.040	0.028	9.0	2	0.127	0.014	9.0
3	0.550	0.044	10.0	3	0.480	0.071	10.0	3	0.624	0.020	10.0
4	0.260	0.038	11.0	4	0.360	0.069	11.0	4	0.206	0.017	11.0
5	0.076	0.023	12.0	5	0.080	0.039	12.0	5	0.032	0.007	12.0
Alleles for the locus 17: DYS448											
No.	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**	**No.**	**Freq.**	**s.d.**	**Allele:**
1	0.008	0.008	16.0	1	0.000	0.000	16.0	1	0.005	0.003	16.0
2	0.015	0.011	17.0	2	0.000	0.000	17.0	2	0.005	0.003	17.0
3	0.008	0.008	18.0	3	0.040	0.028	18.0	3	0.011	0.004	18.0
4	0.275	0.039	19.0	4	0.160	0.052	19.0	4	0.259	0.018	19.0
5	0.580	0.043	20.0	5	0.640	0.069	20.0	5	0.545	0.021	20.0
6	0.053	0.020	21.0	6	0.080	0.039	21.0	6	0.122	0.014	21.0
7	0.000	0.000	22.0	7	0.040	0.028	22.0	7	0.005	0.003	22.0
8	0.015	0.011	23.0	8	0.000	0.000	23.0	8	0.021	0.006	23.0

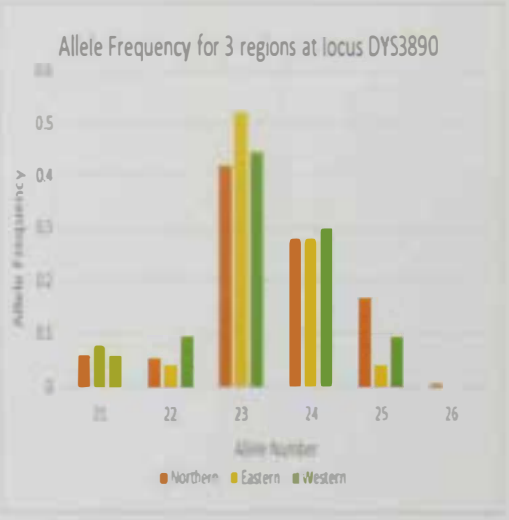
Figure 10(A-Q): Allele frequency of different loci in sub-populations.



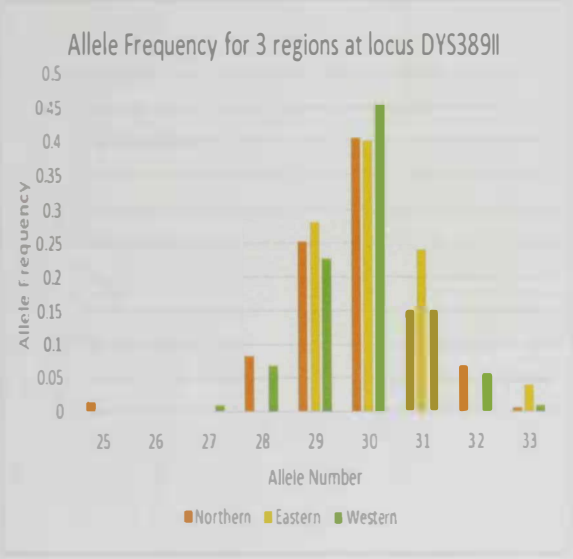
(A)



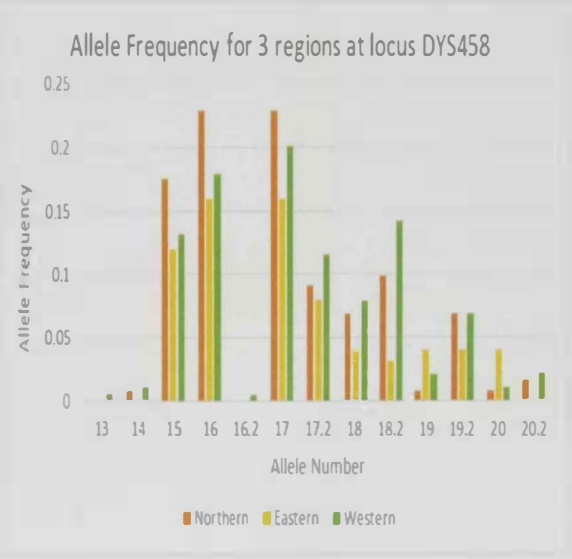
(B)



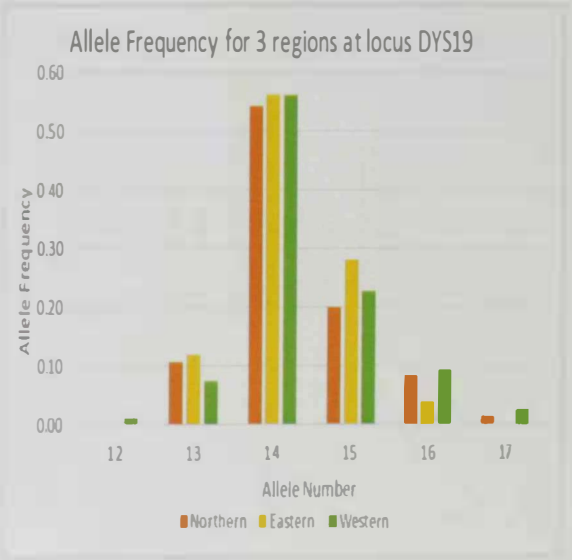
(C)



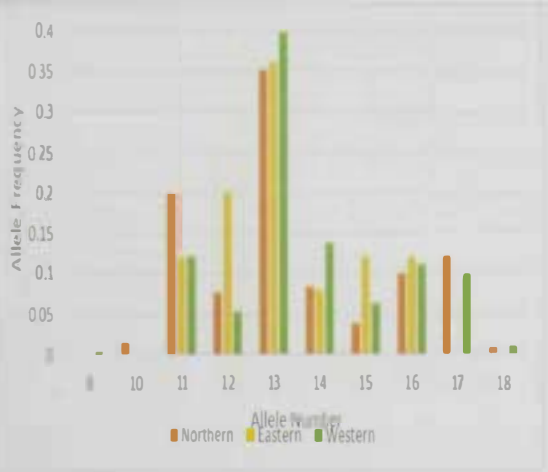
(D)



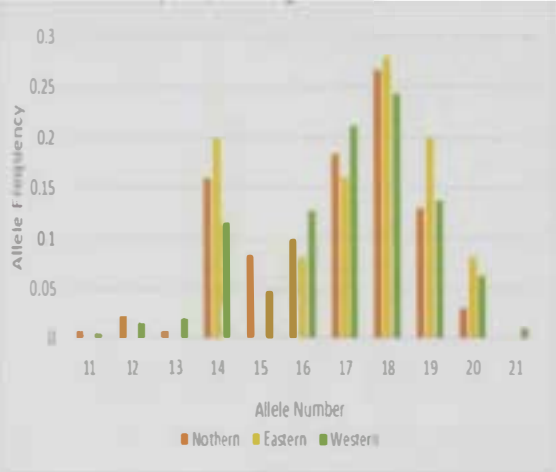
(E)



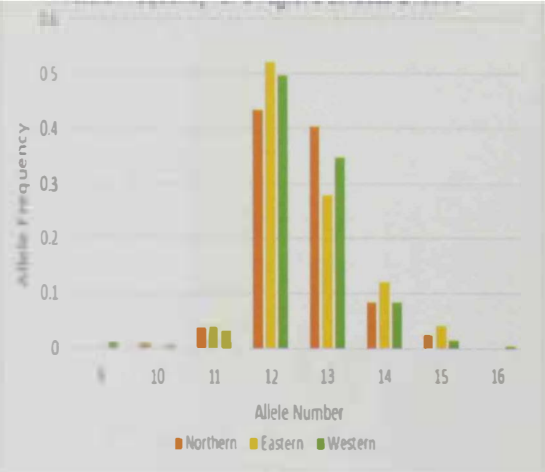
(F)


Allele Frequency
Allele Number
Northern
Eastern
Western

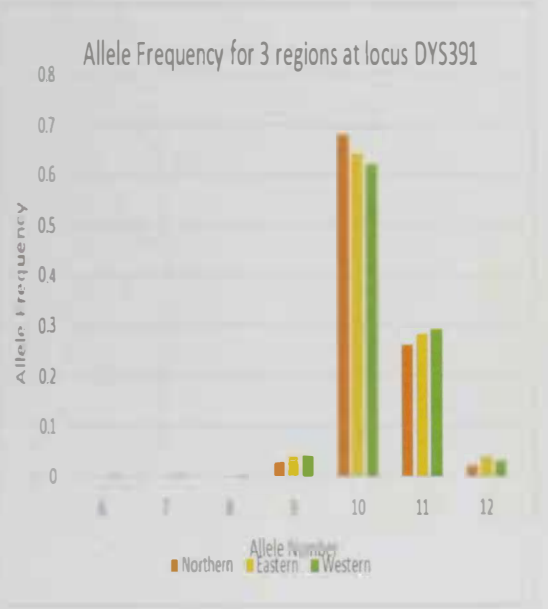
(G)


Allele Frequency
Allele Number
Nothern
Eastern
Western

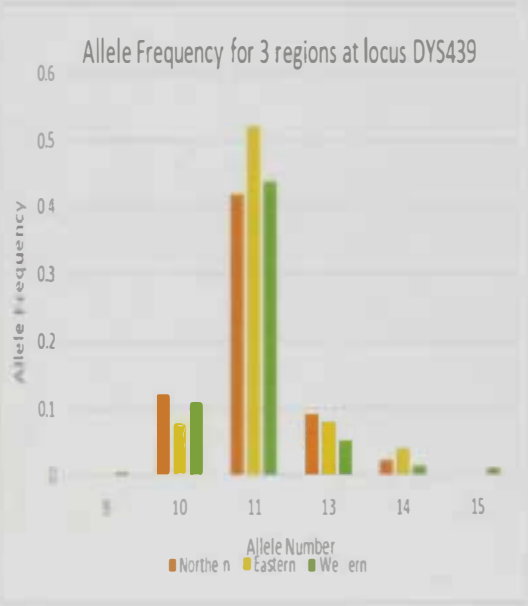
(H)


Allele Frequency
Allele Number
Northern
Eastern
Western

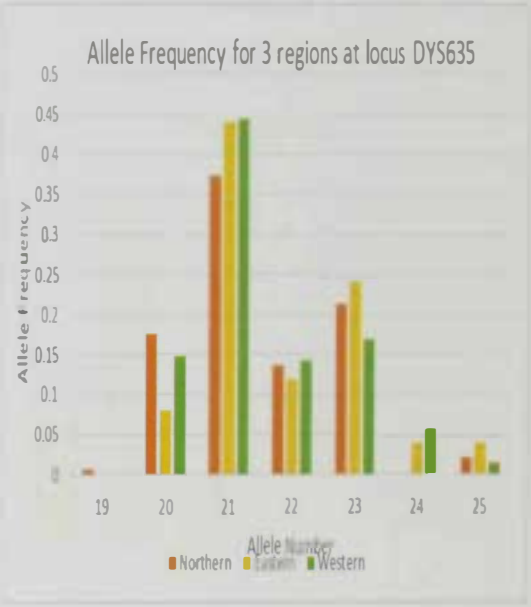
(I)


Allele Frequency for 3 regions at locus DYS391
Allele Frequency
Allele Number
Northern
Eastern
Western

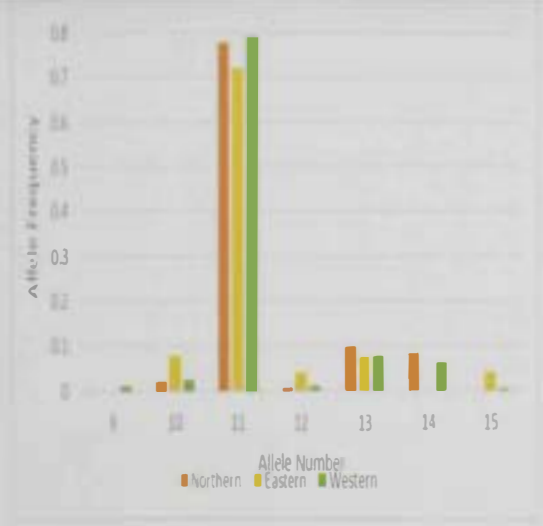
(J)


Allele Frequency for 3 regions at locus DYS439
Allele Frequency
Allele Number
Northern
Eastern
Western

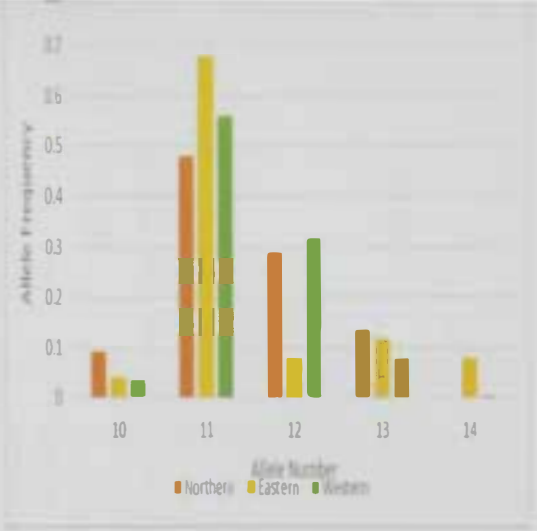
(K)


Allele Frequency for 3 regions at locus DYS635
Allele Frequency
Allele Number
Northern
Western

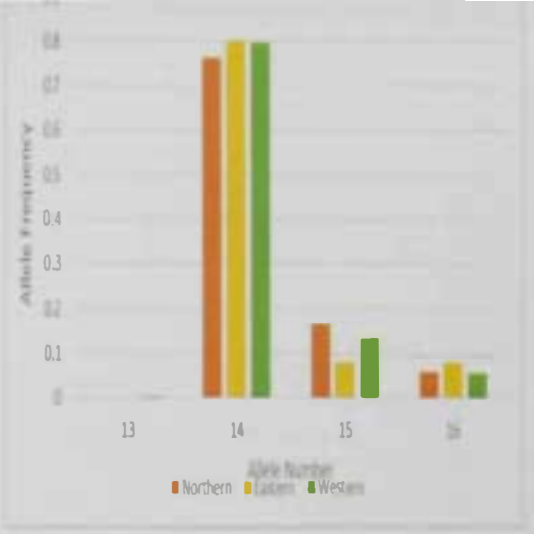
(L)


Allele Frequency
Allele Number
Northern
Eastern
Western

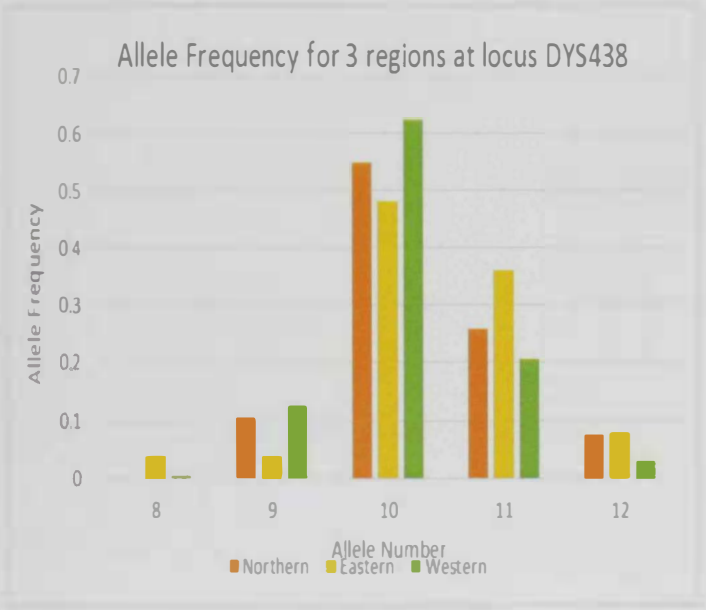
(M)


Allele Frequency
10
11
12
13
14
Allele Number
Northern
Eastern

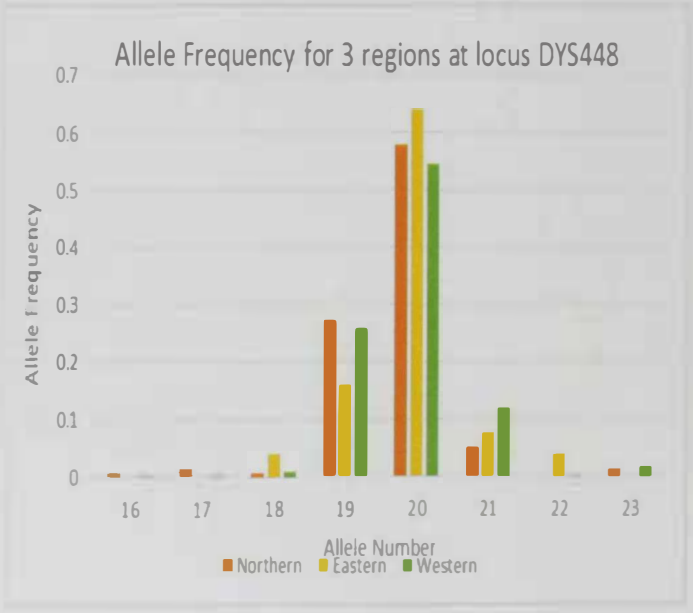
(N)


Allele Frequency
13
14
15
Northern
Eastern
Western

(O)


Allele Frequency for 3 regions at locus DYS438
Allele Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
8
9
10
11
12
Allele Number
Northern
Eastern
Western

(P)


Allele Frequency for 3 regions at locus DYS448
Allele Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
16
17
18
19
20
21
22
23
Allele Number
Northern
Eastern
Western

(Q)

Table 10: Presents a summary of loci with alleles that are not shared between all regions.

Locus	Allele	Northern N=131	Eastern N = 25	Western N=139
DYS456	13	0.061	0	0.032
DYS389I	15	0	0	0.011
DYS390	26	0.008	0	0
DYS389II	25	0.011	0	0
	27	0	0	0.011
	28	0.084	0	0.069
	32	0.069	0	0.058
DYS458	13	0	0	0.005
	14	0.008	0	0.011
	16.2	0	0	0.005
	20.2	0.015	0	0.021
DYS19	12	0	0	0.011
	17	0.015	0	0.026
DYS385-A	8	0	0	0.0053
	10	0.015	0	0
	17	0.122	0	0.101
	18	0	0	0.011
	19	0.008	0	0
DYS385-B	11	0.008	0	0.005
	12	0.023	0	0.016
	13	0.008	0	0.021
	15	0.084	0	0.048
	21	0	0	0.011
DYS393	9	0	0	0.0106
	10	0	0.008	0.005
	16	0	0	0.005
DYS391	6	0	0	0.005
	7	0	0	0.005
	8	0	0	0.005
DYS439	9	0	0	0.005
	15	0	0	0.011
DYS635	24	0	0.040	0.058
	19	0.008	0	0
	25	0.023	0	0.016
DYS392	9	0	0	0.0106
	14	0.024	0	0.63
	15	0	0.040	0.005
GATA_H4	14	0	0.080	0.005
DYS437	13	0	0	0.005
DYS438	8	0	0.040	0.005
DYS448	16	0.008	0	0.005
	17	0.011	0	0.005
	22	0	0.040	0.005

The variation in allele distribution, or alternatively uniqueness, can stem from several factors. There are more common rare alleles that are shared between Northern and Eastern region rather than Western region (see table 10). Alternatively, there are many rare alleles that are only present in Western region and absence from other two regions despite the fact that Northern population size is relatively large. However, one might still argue that these alleles are rare and the population size is not representative of the subpopulation. In fact, most of the 'unique' allele frequencies are <1 %. However, alleles frequencies that are <1% also do show up only in the smaller Eastern subpopulation (n=25). Alternatively, it can be argued that those alleles that are not shared between the regions are the result of new mutations (Kurihara, 2004; Gusmao et al., 2005; for review see Ellegren, 2000). Other explanations for the presence or absence of alleles in any of the subpopulations can stem mainly from population structure. It can reflect a recent gene flow from neighboring regions that contributes minimally to total haplotype frequencies. Furthermore, the gene flow to specific regions are from specific neighboring areas. For example, the Eastern and Northern are coastal areas and it is expected that admixture rate is higher in those regions. Further analysis and experiments at higher molecular levels is required to elucidate the structure of the population with respect to rare alleles.

3.2.3 Variance of allele frequency in UAE population and in sub-population:

Allele variance is a parameter that describes the distribution of an observed population and its mathematical treatment can decipher genetic distance between population and groups within a population. In the recent years, genetic distance

manifested in allele variance rather than allele frequency is increasingly utilized in population genetic studies (Slatkin, 1995; Meirmans, 2006; Excoffier, 2005). We have computed variance of alleles for each locus for the 345 samples. SPSS software was used to calculate the variance for each locus; variance is simply the square root of the standard variation. The variance is obtained for the UAE population as a whole as well as for the three regions (Northern, Eastern and Western) (Figure 11 and 12). It is clear that the variances for each locus is different and reflects the number of alleles and distribution of alleles for the respective locus. However, the comparison of the variance between subpopulations demonstrates clustering for a number of loci while they are spread for other loci (Figure 12).

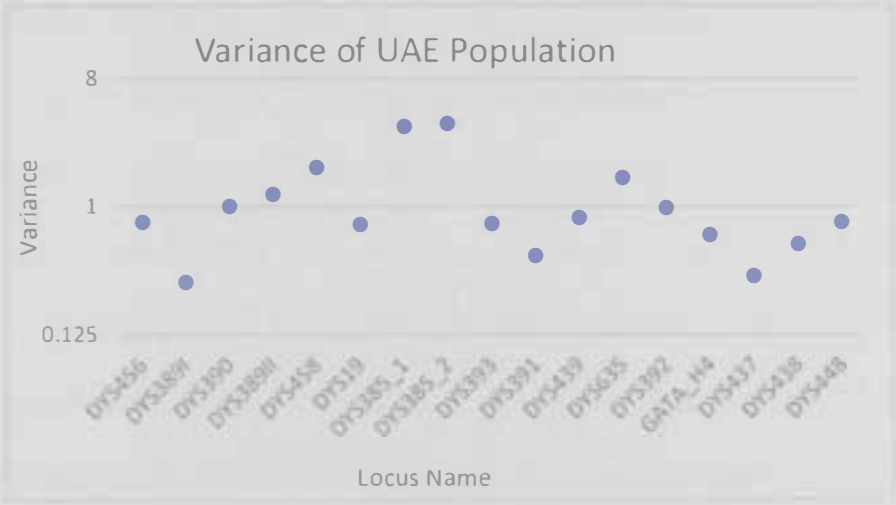


Figure 11: Variance distribution of the UAE population.

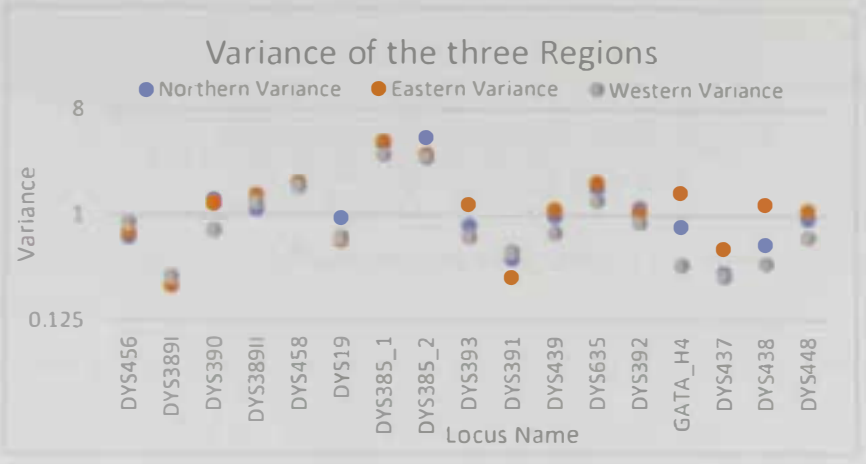


Figure 12: Variance distribution of the three regions.

Since the non-recombining part of Y chromosome is inherited as a unit haplotype and provided that several of the loci's variance are clustered (similar) between subpopulations with no significant difference, then it follows that the variance observed for other loci is probably due to higher susceptibility of those loci to mutation (Gusmao et al., 2005). Normally, STR has a higher rate of mutation than single nucleotides; it is estimated from $1x10^{-2}$- $1X10^{-4}$. These variations depend on the allele itself, position, motif size, sequence composition, and stability of STR (Eckert and Hile 2009; Gusmao et al., 2005). Founder effect and genetic drift are not plausible explanations in the UAE population as the genetic diversity of the 17 STR is high and relatively small members of populations are sharing identical haplotypes (see below). Loci with greater allele variance between the subpopulation are DYS438, DYS437, GATA_H14, DYS19. However, it should be noted that the Western's population allele variance for DYS438, DYS437, GATA_H14, DYS19 loci show departure from Eastern and Northern regions. Although, these results are still inconclusive it warrants further investigation at a higher molecular resolution (SNP). The SNP studies will elucidate whether the variance departure for a specific

region is due to new mutation based on one-step mutation model or to recent gene flow.

3.2.4 Allele Frequency of UAE population compared with other populations:

Predominant alleles observed in the UAE population sample were compared to other predominant alleles in the Arabian Peninsula populations and to surrounding areas using data from other studies performed on the same loci in Y-Chromosome. Comparison of allele frequencies with other populations in the region of the Arabian Peninsula and its neighboring countries (Soares et al., 2008; Hedman et al., 2004; Fadhlaoui et al., 2012; Ambrosio et al., 2012; Imad et al., 2013; Yunis et al., 2013; Hallenberg et al., 2005; Rosa et al., 2006; Rahman et al., 2012; Dönbak et al., 2006; Turrina et al., 2006; Chang et al., 2007; Abdin et al., 2003) show similar prevalence of predominate alleles across the loci. It can be seen that the UAE population share most of its predominant alleles with Turkey (shares 13 loci), Tunisia (shares 10 loci), Syria (shares 9 loci), Iraq and India (they share 9 loci). In general, they share more than 50% of their predominant alleles with the UAE population (see table 11). However, it should be noted that each population retain their particularities of allele variance and frequencies that can place them at particular genetic distance.

Table 11: UAE population allele frequency compared with other populations. Alleles shared between UAE and other populations are highlighted.

Population	UAE N=345	Guinea Bissau N=215	Iraq N=400	Turkey N= 86	Syria N=113	Somalia N=201	Spain N=347	Tunisia N=218	Brazil N=412	Finland N=400	Pakistan N=71	Colombia N=173	Italy N=155	India N=106
DYS456	15	-	15	15	-	-	14	15	-	-	15	-	-	15
DYS389I	13	13	14	13	13	13	13	13	13	14	13	13	13	13
DYS390	23	21	23	23	23	25	24	25	24	24	24	24	24	22
DYS389II	30	30	29	29	30	31	16	31	29	30	30	29	29	30
DYS458	17.16*	-	15	16	-	-	16	17	-	-	15	-	-	16
DYS19	14	14	14	14	14	11	14	13	14	14	15	14	14	15
DYS385-A	13	-	13	10	16	27	15	15	-	11	11	-	12	11
DYS385-B	18	-	14	11	18	28	18	18	-	13	14	-	-	14
DYS393	12	13	14	12	12	13	13	13	13	14	13	13	13	13
DYS391	10	10	10	10	10	10	11	9	10	11	11	11	10	10
DYS439	11	13	10	11	12	11	12	10	-	10	10	14	12	11
DYS635	21	-	24	21	-	-	22	21	-	-	21	24	-	-
DYS392	11	11	11	11	11	13	14	11	13	14	11	13	11	11
GATA_H4	11	-	12	11	-	-	12	11	-	11	12	29	-	12
DYS437	14	14	14	13	14	14	15	14	-	14	14	15	15	14
DYS438	10	11	10	10	9	11	12	10	-	10	11	13	-	11
DYS448	20	-	19	20	-	-	19	20	-	-	20	-	-	19
Number of Shared loci	17	6	8	13	9	4	4	10	3	5	7	2	4	8

*Two predominant allele in the UAE population.

3.3 Y STR haplotype of UAE population:

Briefly, the word "haplotype" describes a genetic unit or combination of alleles at adjacent locations or loci on the chromosome that are inherited together from a single parent. Autosomal haplotypes are prone to recombination and therefore the erosion of its gametic phasing (Hartl, 1997). On the other hand, an organism with heterogametes, one of the chromosomes that is mostly non-recombining is inherited as one haplotype (Kayser et al., 1997). Haplotype can appear in many different ways such as, one locus, several loci or an entire chromosome depending on recombination number in a set of loci. There are three primary reasons for considering the haplotype organization of a variation. First is that, the unit of biological function, the protein-coding gene, produces proteins whose sequences correspond to maternal and paternal haplotypes. Second, the variation in a population is in fact, structured into haplotypes that are likely transmitted as a unit. Lastly, regardless of the population genetic reasons, haplotypes serve to reduce the dimensionality of the problem of testing association, and so they may increase the power of those tests (Clark, 2004).

In this study, 345 haplotypes each with 17 alleles were analyzed using Arlequin software (Excoffier, 2005). The analysis measures the frequency of alleles and genetic variations between populations, within a population, and among groups.

3.3.1 Haplotype Frequency:

Haplotypes on non-psuedoautosomal region of Y- chromosome passes from one generation to another intact except where mutation events have taken place. This study measures similarities or dissimilarities between haplotypes to describe the

genetic structure of UAE population. Although in our study, the population shows relatively diverse haplotype, there are a number of haplotype shared at least between two individuals who are not related (see table 12).

Table 12: Haplotype frequency for shared haplotype in the UAE population. (fi represent number of count samples times squared frequency).

Haplotypes Frequencies							
Shared Haplotype	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Freq.	0.002899	0.005797	0.008696	0.011594	0.014493	0.017391	-
(Freq.)^2	$8.4*10^{-06}$	$3.36*10^{-05}$	7.5610^{-05}	0.000134	0.00021	0.000302	-
count	271	22	5	1	1	1	301
fi	0.002277	0.000739	0.000378	0.000134	0.00021	0.000302	0.004041
N	271	44	15	4	5	6	345

3.3.2 Haplotype Diversity:

The presence of identical haplotypes warrant the investigation of haplotype diversity. Haplotype diversity is a measure of the uniqueness of a particular haplotype in a given population. This measure of gene diversity is analogous to hetrozygosity at a single locus. The haplotype diversity (H) in our population was calculated using the following formula:

$$H = \frac{N}{N-1}(1 - \sum_i x_i^2)$$

Where x_i is the relative haplotype frequency of each haplotype in the sample and N is the number of sample (Nei et al., 1987). Haplotype diversity in the UAE population is equal to 99.885 % (see table 14 in section 3.4), which means that there is a large genetic diversity in UAE population. However, we have also compared

genetic distance between the subpopulation in the three regions. Excel-based (Microsoft®) software GENALEX 6 was used to compute Nie's pairwise genetic matrix (Peakall and Smouse, 2006). The results of pairwise population Nei's genetic distance analysis show a relatively small difference between populations. The pairwise genetic distance computation is based on stepwise mutation models (Nei, 1978). It should be noted that region 2 (Eastern) demonstrates highest distance when compared to Western region and not as large of a genetic distance when compared to Northern (see table 13). This can be attributed to sample size but geographical location of the latter two regions (Eastern and Northern) may account for the genetic similarities, as these two regions are coastal regions and may have shared the same migratory paths. Nie's genetic distance (Takezak and Nei, 1996; Nei, 1978; Peakall Smouse, 2006), however, measures the summation of all genetic loci examined for the population under study and it is prone to biases for two fundamental reasons: 1) the sample size and 2) the number of genetic loci examined. Although, this study has experimented with genetic distance within the UAE population, the current information must be cautiously interpreted but warrant further investigation at higher molecular resolution (i.e. SNP, copy number variation). It is also important to appreciate that the geographical location of different regions in the UAE provided different resources and attracted different interests for immigrants. This is especially true for the periods before 1970's distance (Nei, 1978).

Table 13: Represent the results Pairwise Population Matrix of Nei Genetic Distance for Northern (1), Eastern (2), and Western (3)

Pairwise Population Matrix of Nei Genetic Distance and Identity					
Pop1	Pop2	Nei genetic Distance	Nei genetic identity	#Pop1	#Pop2
1	2	0.069	0.934	131	25
1	3	0.043	0.958	131	189
2	3	0.139	0.870	25	189

3.4 Discrimination Capacity:

By definition, Discrimination Capacity (DC) means that the number of haplotypes observed only once in the population, where designated as "Unique haplotype". The DC calculated for the 17 loci studied as a percentage of unique haplotype by dividing the number of unique haplotype over the total number of the haplotypes. From table 13 the percentage of DC is equal to 90.03 which indicates that the UAE population is comprised of a relatively large number of unique haplotypes. In this study none of the shared haplotypes comes from first degree relatives (fathers/son brother, paternal cousins). This information has a significance impact in forensic analysis and other loci on Y chromosome should be used in conjunction with the 17 loci utilized. Furthermore, the DC indicates that there are common haplotypes in the UAE population that is probably contributed by sharing recent ancestors. The UAE population demography, social structure, and culture strengthen why DC is not approaching 100%.

Table 14: shows the percentage of Discrimination Capacity and Haplotype Diversity in UAE population

Forensic Parameter of Y Filer in UAE population	
Samples in Population	345
Number of haplotypes	301
Number of unique haplotypes	271
Discrimination Capacity (%)	90.03
Haplotype Diversity (%)	99.885

3.5 Shared haplotype in UAE population:

Common haplotype among closely located populations may imply common paternal ancestry (Furedi et al., 1999). The finding of shared identical haplotype in a population has a significant connotation on forensic investigations with regard to probability of identity between different populations. The UAE population investigated in this study is comprised of 301 haplotypes. Of the total haplotype (n=301), 271 haplotypes are unique while 5 haplotypes are shared at least between two individuals (see figure 13). While the majority of shared haplotypes occurs between two individuals, there are three cases where four or more individuals are sharing identical haplotype. Moreover, there are three different haplotypes shared by five individuals (see Table 15 (A-E). In most cases these individuals are not from the same clan. However, in one instance, there are three individuals with the same tribe name who are sharing the same haplotype; upon further investigation, the three individuals do not share father or grandfather names.

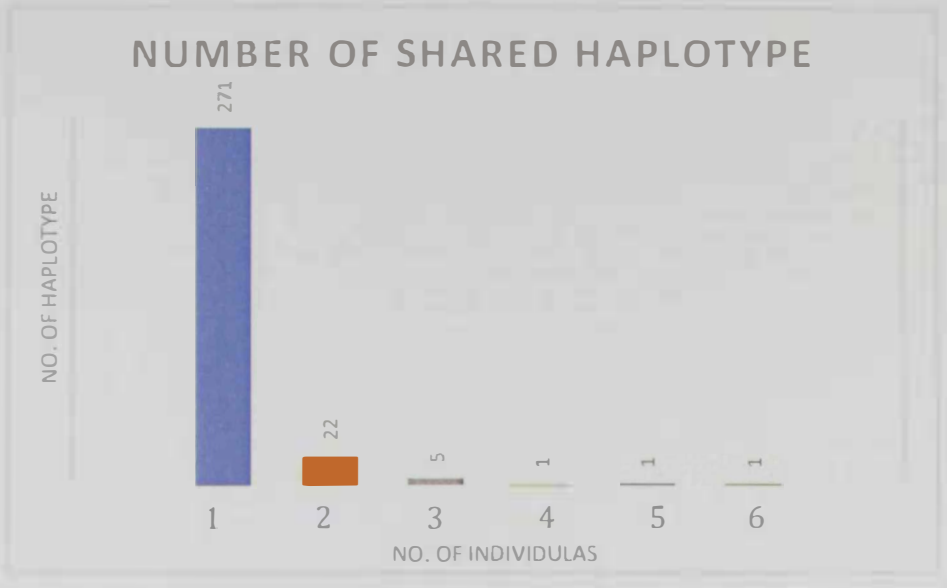


Figure 13: Total number of Unique and Shared Haplotype

Table 15(A-E): Presents the haplotypes that are shared in the population. A) Haplotypes shared between two individuals; B) Haplotypes shared between three of Individuals; C) Haplotypes shared between four of Individuals; D) Haplotypes shared between five of Individuals. E) Haplotypes shared between six Individuals. The order of alleles in the tables are DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385A, DYS385B, DYS438, DYS439, DYS437, DYS448, DYS456, DYS458, DYS635 and GATA-H4.

A

No.	Freq.	S.D	Haplotype																
26	0.005797	0.004093	16	13	23	31	172	14	13	17	12	10	11	20	11	11	14	10	20
42	0.005797	0.004093	15	13	23	30	172	14	13	16	12	10	11	20	11	11	14	10	20
44	0.005797	0.004093	15	13	23	31	15	16	11	15	13	10	10	23	11	13	14	11	20
46	0.005797	0.004093	14	12	24	28	18	15	16	20	12	9	12	21	11	12	14	9	19
57	0.005797	0.004093	14	13	23	30	182	14	13	19	12	11	11	21	11	11	14	10	20
59	0.005797	0.004093	16	13	23	29	17	14	12	14	12	11	13	23	13	13	14	12	19
63	0.005797	0.004093	16	13	25	30	15	13	16	17	13	10	11	21	11	10	14	10	20
86	0.005797	0.004093	14	13	23	29	182	14	13	18	12	10	11	21	11	10	14	10	20
89	0.005797	0.004093	14	13	23	30	182	14	13	20	12	11	11	22	11	11	14	10	20
91	0.005797	0.004093	14	13	23	30	182	14	13	18	12	10	11	21	11	11	14	10	20
100	0.005797	0.004093	14	13	23	30	182	14	13	18	12	11	11	21	11	11	14	10	19
127	0.005797	0.004093	15	13	24	30	15	14	15	17	13	11	12	22	11	12	14	10	20
139	0.005797	0.004093	16	14	24	29	17	14	11	14	12	10	15	23	14	12	15	12	19
145	0.005797	0.004093	14	13	23	30	182	14	13	18	12	11	11	22	11	11	14	10	19
160	0.005797	0.004093	15	14	25	32	16	15	11	14	13	10	10	23	11	12	14	11	20
161	0.005797	0.004093	16	13	24	30	16	14	17	17	13	10	12	20	11	12	14	10	20
178	0.005797	0.004093	16	13	24	30	17	14	17	18	13	10	12	20	11	12	14	10	20
184	0.005797	0.004093	14	14	23	31	182	14	13	18	12	11	11	21	11	11	14	10	20
190	0.005797	0.004093	15	13	25	29	16	15	11	15	14	11	10	24	11	13	14	11	19
217	0.005797	0.004093	13	14	24	32	17	15	11	12	13	10	12	?	11	12	14	10	23
280	0.005797	0.004093	15	14	23	31	17	14	14	16	13	10	11	21	13	11	14	10	19
7	0.005797	0.004093	17	13	24	30	15	13	15	17	13	10	11	22	11	11	14	10	20

B

No.	Freq.	S.D	Haplotype:																
183	0.0087	0.0050	14	13	23	30	192	14	13	19	12	11	11	21	11	11	14	10	20
25	0.0087	0.0050	14	13	23	30	182	14	13	19	12	11	11	22	11	11	14	10	20
118	0.0087	0.0050	14	13	23	30	192	14	13	18	12	12	12	21	11	11	14	10	20
72	0.0087	0.0050	15	13	23	30	172	14	13	17	12	10	11	20	11	11	14	10	20
9	0.0087	0.0050	14	13	23	30	182	14	13	18	12	11	11	21	11	11	14	10	20

C

No.	Freq.	S.D	Haplotype:																
14	0.0116	0.0058	16	13	24	30	16	14	17	18	13	10	12	20	11	12	14	10	20

D

No.	Freq.	S.D	Haplotype:																
85	0.0145	0.0064	15	13	25	30	15	13	16	17	13	10	11	21	11	10	14	10	20

E

No.	Freq.	S.D	Haplotype:																
12	0.0174	0.007048	15	13	23	30	172	15	13	18	12	10	11	20	11	11	14	10	20

This is likely due to the sharing of most common recent ancestors. However, in the majority of cases, the tribes name does not predict a specific haplotype. This is not surprising as tribal names, although it stems from a common founder; it is given to individuals that do not have direct blood relationship. These results have significant implications on forensic science practices. These results call for more thorough analysis that should include SNP studies alongside the 17 Y chromosome STR analysis. We also have found that approximately 50% of the UAE population share between 6 to 7 loci the result, which should restrict the usage of minimal panel of STR (see Figure 14). The data was analyzed by GENALEX 6 which is built as Excel (Microsot®) add-in program (Peakall and Smouse, 2006). Figure 14 presents the number of matching haplotypes versus unique haplotype by locus.

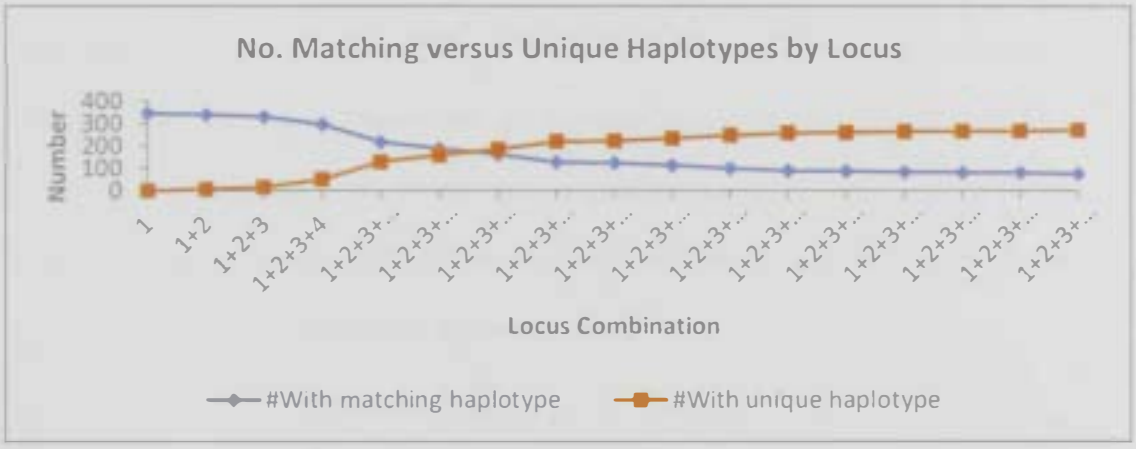


Figure 14: Matching versus Unique Haplotypes by Locus.

3.5.0 Shared haplotype of UAE population compared with other populations:

In our population, 301 different haplotypes were identified, 271 of which were unique. The most frequent haplotype was found in 22 instances. The second most frequent haplotype were 4, 5, 6, which they shared only one haplotype. In addition, by comparison with other populations, we found that in Libya (n= 176 individuals), they have a total of 142 different haplotypes and 124 was unique (Triki et al., 2013). However, in China (n=112 individuals) a total of 99 haplotypes were found and 88 of them were unique haplotypes (Tie et al., 2003). In addition, Syria with 113 individuals, they observed that there are 108 different haplotypes and 104 haplotypes were found to be unique (Abdin et al., 2003). Moreover, in the Kuwaiti population a total of 101 different haplotypes, among which 78 where unique, eight haplotypes were shared between two individuals, three individuals shared four haplotypes, four individuals shared only one haplotype and the most frequent haplotype was shared by 7 individuals (Soumaya et al., 2010). In Tunisa, they found that the total number of haplotypes were 154, of which 127 were unique, the most common haplotype was represented by 14 individuals (Fadhlaoui-Zid et al., 2012).

3.5 Gene diversity for 17 polymorphic loci in UAE population and in sub-population:

Gene diversity for all 345 samples were calculated using the following formula:

$$H = \frac{N}{N-1}(1 - \sum_i x_i^2)$$

Where x_i is the frequency of each allele in the locus and N is the number of allele presented in the locus (Nei et al., 1987). According to measures of genetic diversity, the highest diversity were observed at locus DYS458 = 0.9, DYS385-A = 0.9 and DYS385-B = 0.9. Therefore, they should be considered as the most variable and most informative markers for forensic testing. While, loci with the lower diversity are the least informative loci (DYS392 which equal 0.437; see table 16).

Another parameter that can be calculated is the total genetic diversity and we calculated it by dividing the number of observed haplotype over the total samples which is equal in this study 87%. In the sub-population, there are varieties of the highest and lowest of alleles frequencies. For example, in Northern region, there are nine loci with highest and two loci with lowest diversity. While in Eastern region, the highest gene diversity is in DYS389II with a value of 1.00 and the lowest involves the DYS437 locus; moreover in Western region, the highest gene diversity is associated with five different locus (See table 16). The trend of gene diversity in the sub-population is not different from the whole UAE population. However, the differential diversity per locus reinforces a stepwise mutation model in which the memory of the previous event of mutation is maintained (Valdes, 1993) as opposed to infinite mutation model (Hudson, 2002).

Table 16: Genetic diversity for 17 polymorphic loci in UAE population and in sub-population.

Locus	UAE population	Northern	Eastern	Western
DYS456	0.8	0.9	0.9	0.8
DYS3891	0.6	0.5	0.7	0.6
DYS390	0.8	0.9	0.8	0.9
DYS3891I	0.8	0.9	1.0	0.8
DYS458	0.9	0.9	1.0	0.9
DYS19	0.7	0.8	0.8	0.7
DYS385_A	0.9	0.9	1.0	0.9
DYS385_B	0.9	0.9	1.0	0.9
DYS393	0.7	0.7	0.8	0.7
DYS391	0.6	0.6	0.7	0.6
DYS439	0.8	0.9	0.8	0.8
DYS635	0.8	0.9	0.9	0.9
DYS392	0.4	0.5	0.6	0.3
GATA_H4	0.8	0.9	0.6	0.7
DYS437	0.5	0.6	0.5	0.5
DYS438	0.7	0.8	0.8	0.7
DYS448	0.7	0.7	0.7	0.7

3.6 AMOVA results in UAE population:

Analysis of Molecular Variance (AMOVA) was calculated in Arlequin software over two sources of variation among the population and within population as shown in table 17. AMOVA measures haplotype diversity rather than just allele frequencies that provides an opportunity to measure the difference in haplotypes in a pairwise manner. Furthermore, the analysis accommodates and relies less on different types of assumption about the evolution of genetic models. It is clear from the AMOVA that the majority >99.5% of the variations within a population rather than between populations. This reinforces our earlier discussion that Nie's (1987) genetic distance calculations suffered from: 1) number of loci, 2), population size (especially for Eastern regions), and 3) the assumptions that are an intrinsic part of mathematical

treatment. In addition the average F-Statistics over all loci Fixation Indices (Fst = 0.00280). Although Fst calculation depends on the allele frequencies, it is in concordance with lack of hierarchy in the UAE sub-population.

Table 17: AMOVA design and results (average over 17 loci):

Source of Variation	Sum of squares	Variance components	Percentage Variation
Among populations	15.065	0.015	0.280
Within populations	3978.801	5.341	99.720
Total	3993.866	5.356	100

However, there are appreciable genetic diversity in the UAE population as a whole that reflect high rate of admixture different geographical regions. AMOVA shows (table 18) that UAE population is closest to Yemen, Kuwait, Iraq and Iran. Gene flow accounts for these similarities and reflects the history of the UAE which is in accordance with the history of Arabian Peninsula's history.

Table 18: AMOVA results in other population:

Population	Afghanistan	Iran	Iraq	Kuwait	Yemen	UAE
Afghanistan	0					
Iran	0.0026	0				
Iraq	0.0027	0.0002	0			
Kuwait	0.0039	0.0014	0.0014	0		
Yemen	0.0031	0.0006	0.0006	0.0018	0	
UAE	0.0031	0.0008	0.0008	0.0019	0.001	0

CHAPTER IV: CONCLUSION

Conclusion:

In this study, buccal swap samples were obtained from 345 unrelated individuals from the United Arab Emirates population. All of the DNA samples were performed in multiplex fashion using AmpFLSTR® Yfiler® PCR Amplification Kit (Life Technologies) to co-amplify 17 Short tandem repeats (STRs) loci. Experimental variation was reduced with the use of highly automated system called 3500XL Genetic Analyzer (Life Technologies). The UAE population was divided into three regions according to geographical region and proximity, namely, Northern, Eastern and Western regions. In the current study, we have estimated allele frequency, haplotype frequencies, variances, and used analysis of molecular variance (AMOVA) to stratify the population.

The 17 Y chromosome STRs analyzed in this study proved to be highly informative markers, with high values for gene diversity with relatively high discriminatory capacity. Thus, based on this study, the application 17 Y chromosome STRs analysis for forensic and paternity analysis proves to be useful in the UAE. However, further, genetic analysis such as SNP or copy number variation such as Alu sequence is required for definitive confirmation in forensics investigation or deep ancestry studies.

The allelic profile, frequency, and distribution for each locus in the three subpopulations of the UAE follow similar patterns. The allele frequency among the three subpopulations, however, shows small fluctuations especially for specific loci and that could be because of the number of samples. Although, our study demonstrates that there are unique alleles to a subpopulation or not shared by all regions, the profile of the allele frequency and distribution are very similar. We

strongly believe that a larger population size from different emirates will ultimately identify the uniqueness of these alleles in the respective subpopulations.

Although in our study, the population shows relatively diverse haplotype, there are 271 unique haplotype. Twenty-two of the haplotypes are shared between two individuals. There are other haplotypes that are shared between at least two individuals. Haplotype diversity in the UAE population is equal to 99.885 %, which means that there is a large genetic diversity in UAE population. On the other hand, the percentage of discrimination capacity is equal to 90.03%, which indicates that UAE population consists of individuals who share identical haplotype which are not related as judged by the names.

This study is very important and it contributes to other studies around the Gulf Area to study the genetic diversity of population. However, this study calls for more extensive genetic study of the region in order to elaborate on genetic distance between population, disease-associated haplotypes, archeological, and historical studies.

BIBLIOGRAPHY

Bibliography:

Abdin, Louai, et al. "Analysis of 13 Y-chromosomal STRs in an Arab population sample from Syria." International Congress Series. Vol. 1239. Elsevier, 2003.

Abu-Amero. Kh., Hellani. A., González. A., Larruga. J., Cabrera.V., Underhill. P. (2009) . Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. BMC Genetics, 10.1186(1471-2156),10-59.

Ali S, Hasnain SE. Genomic Of The Human Y-Chromosome: association with male infertility. Gene 2003:321:25-37.

Al-Sayegh F.(1998). UAE from tribe to country. Al_Khaleej center for books, Dubai, UAE. (in Arabic).

Ambrosio, Beatriz, et al. "Y-STR genetic diversity in autochthonous Andalusians from Huelva and Granada provinces (Spain)." Forensic Science International: Genetics 6.2 (2012): e66-e71.

Ballantyne J., Sensabaugh G. and Witkowski J., (1989). DNA Technology and Forensic Science. 32 Banbury report, USA.

Balloux, F. and Lugon-Moulin, N. (2002), The estimation of population differentiation with microsatellite markers. Molecular Ecology, 11: 155–165. doi: 10.1046/j.0962-1083.2001.01436.

Bao W, Pandya A, Zerja T, Xu J, Shu Q, Du R, Yang H, Tyler-Smith C.(2000). MSY2: a slowly evolving minsatellite on the human Y chromosome which provides a useful polymorhoric marker in Chinese populations.Gene. 244:29-33.

Becker WM, Kleinsmith LJ, Hardin J. The world of the cell, 6th ed., San Francisco: Benjamin Cummings Publisher; 2006. 796 p.

Becker, W.M., Reece, J.B., Ponie, M.F. 1995, the world of the cell.3rd ed. The Benjamin/Cummings Publishing Company.

Bortolini, M. C., Salzano, F. M., Thomas, M. G., Stuart, S., Nasanen, S. P. K., Bau, C. H. D., Hutz, M. H., Layrisse, L., Erler, M. L. P., Tsuneto, L. Y. Y., Hill, H., Hurtado, A. M., de-Guerra, D. C., Torres, M. M., Groot , H., Michalski,, R., Nymadawa, P., Bedoya, G., Bradma, N., Labuda, D., Linares, A. R. (2003). Y-Chromosome Evidence for Differing Ancient Demographic Histories in the Americas. M. J. Hum. Genet. 73, 524–539.

Bosch, E., et al. "Y chromosome STR haplotypes in four populations from northwest Africa." International journal of legal medicine 114.1-2 (2000): 36-40).

Brinkmann, Bernd, et al. "Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat." The American Journal of Human Genetics 62.6 (1998): 1408-1415.

Butler J M. (2011). Advanced Topic in Forensic DNA Typing, 3rd Edition. Elsevier Academic Press, London.

Cadenas, Alicia M., et al. "Y-chromosome diversity characterizes the Gulf of Oman." European Journal of Human Genetics 16.3 (2007): 374-386.

Carvalho-Silva, D.R., Santos, F.R., Rocha,J., Pena, S.D.H. (2001). The Phylogeography of Brazilian Y-Chromosome Lineages. Am. J. Hum. Genet, 68, 281–286.

Chang, Yuet Meng, et al. "A distinct Y-STR haplotype for Amelogenin negative males characterized by a large deletion." Forensic science international 166.2 (2007): 115-120.

Clark, Andrew G. "The role of haplotypes in candidate gene studies." Genetic epidemiology 27.4 (2004): 321-333.

Collins, Patrick J., et al. "Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR Identifiler PCR Amplification Kit." Journal of forensic sciences 49.6 (2004): 1265-1277.

"Censuses of the UAE." Uae statistics, 25/12/2010 http://www.uaestatistics.gov.ae/EnglishHome/tabid/96/Default.aspx, 10th, May 2014.

Daniel D. Chiras. Human Biology, 6th ed., Canada: Jones and Bartlett Publishers Company; 2008. 496 p.

David Bainbridge, 'The X in Sex: How the X Chromosome Controls Our Lives, pages 3-5, 13, Harvard University Press, 2003 ISBN 0674016211).

Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. European Journal of Human Genetics,17, 91–99.

Dönbak, Lale, et al. "Y-STR haplotypes in populations from the Eastern Mediterranean region of Turkey." International journal of legal medicine 120.6 (2006): 395-396.

Dupuy, B. Myhre, et al. "Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci." Human mutation 23.2 (2004): 117-124.

Eckert, K. A. and S. E. Hile (2009). "Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome." Mol Carcinog 48(4): 379-388.

Ellegren, Hans. "Microsatellite mutations in the germline: implications for evolutionary inference." Trends in genetics 16.12 (2000): 551-558.

Excoffier, L. G. Laval, and S. Schneider (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Excoffier, Laurent, Guillaume Laval, and Stefan Schneider. "Arlequin (version 3.0): an integrated software package for population genetics data analysis." Evolutionary bioinformatics online 1 (2005): 47.

Fadhlaoui-Zid, Karima, et al. "Sousse, Tunisia: Tumultuous history and high Y-STR diversity." Electrophoresis 33.23 (2012): 3555-3563.

Gaetano, C. D., Cerutti, N., Crobu, F., Robino, D., Inturri, S., Gino, S., Guarrera, S., Underhill, P. U., King, R. H., Romano, V., Cali, F., Gasparini, M., Matullo, G., Salerno, A., Torre, C., Piazza, A. (2009).

Gill, P., Kimpton, C., Urquhart, A., Oldroyd, N., Millican, E., Watson, S. and Downes, T. (1995). Automated short tandem repeat (STR) analysis in forensic casework – a strategy for the future. Electrophoresis 16: 1543-1552.

Gusmao, L., et al. "Mutation rates at Y chromosome specific microsatellites." Human mutation 26.6 (2005): 520-528.

Hald, Anders (1998). A History of Mathematical Statistics. New York: Wiley. ISBN 0-471-17912-4.

Haley, Hunter-Zinck., Shaila, Musharoff., Jacqueline, Salit., Khalid, A. Al-Ali., Lotfi, Chouchane.,Abeer, Gohar.,et al. (2010): Population Genetic Structure of the People of Qatar. The American Journal of Human Genetics 87, 17–25.

Hallenberg, Charlotte, et al. "Y-chromosome STR haplotypes in Somalis." Forensic science international 151.2 (2005): 317-321.

Harris, P., Boyd,E., Young, B.D., Ferguson –Smithm M.A., 1986. Determination of the DNA content of human chromosomes by flow cytometry. Cytogenet. Cell Genet.41, 14-21.

Hart D. L. 2011. Essential Genetics a Genetics Perspective, 5th ed. MA: Jones and Bartlett Publishers.

Hartl, Daniel L., and Andrew G. Clark. Principles of population genetics. Vol. 116. Sunderland: Sinauer associates, 1997.

Hedman, Minttu, et al. "Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages." Forensic science international 142.1 (2004): 37-43.

Heuertz, Myriam, et al. "Nuclear microsatellites reveal contrasting patterns of genetic structure between western and southeastern European populations of the common ash (Fraxinus excelsior L.)." Evolution 58.5 (2004): 976-988.

Holmlund, G., et al. "Y-chromosome STR haplotypes in a Swedish population." International Congress Series. Vol. 1239. Elsevier, 2003.

Edwards, A., Hammond, H., Jin, L., Caskey, C. and Chakraborty, R. (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12, 241-253.

http://www.nature.com/scitable/definition/haplotype-haplotypes-142, 14th, May 2014.

Hudson, Richard R. "Generating samples under a Wright–Fisher neutral model of genetic variation." Bioinformatics 18.2 (2002): 337-338.

Imad, H., et al. "Genetic variation of 17 Y-chromosomal short tandem repeats (STRs) loci from unrelated individuals in Iraq." International Journal of Biotechnology and Molecular Biology Research 4.8 (2013): 119-129.

Jarne, Philippe, and Pierre JL Lagoda. "Microsatellites, from molecules to populations and back." Trends in Ecology & Evolution 11.10 (1996): 424-429.

Jobling MA, Tyler-Smith C. (2003). The human Y chromosome: an evolutionary marker comes of age. Nat. Rev. Genet, 4:598–612.

Jobling MA, Williams G, Schiebel K, Pandya A, McElreavey K, Salas L, Rappold GA, Affara NA, Tyler-Smith C. (1998)A selective Difference between human Y-chromosomeal DNA haplotypes. Current Biology. 8:1391-1394.

Krausz C, Quintana-Murci L, Forti G. (2004). Y chromosome polymorphisms in medicine. Ann. Med, 36, 573–583.

Krawczak, M and Schmidtke J. DNA fingerprinting, Oxford UK: Bios Scientific Publishers, 1994, Vol. 23, 69-107 P.

Kurihara, Rina, et al. "Mutations in 14 Y-STR loci among Japanese father-son haplotypes." International journal of legal medicine 118.3 (2004): 125-131.

Lodish, Molecular Cell Biology, 6th ed., W.H.Freeman and Company. 2007. ISBN 0-7167-7601-4.

Manuelidis L., (1978). Complex and simple sequences in human repeated DNAs. Chromosome 66, 1-21.

Meirmans, Patrick G. "Using the AMOVA framework to estimate a standardized genetic differentiation measure." Evolution 60.11 (2006): 2399-2402.

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics, 142, 1061–1064.

Milkos G., and John B., (1979). Hetrochromatin and satellite DNA in man: properties and prospects. Am. J. Hum Genet. 31, 264-280.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 76: 379-390.

Peakall, R. O. D., and Peter E. Smouse. "GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research." Molecular ecology notes 6.1 (2006): 288-295.

Pestoni, C., et al. "Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain)." International journal of legal medicine 112.1 (1998): 15-21.

Rablen, Matthew D., and Andrew J. Oswald. "Mortality and immortality: The Nobel Prize as an experiment into the effect of status upon longevity." Journal of health economics 27.6 (2008): 1462-1471.

Rahman, Z., and T. Husnain. "Y-Chromosomal STR Haplotype Profiling in Yousafzai's living in Swat Valley Pakistan. 2008".

Robino, C., et al. "Analysis of Y-chromosomal SNP haplogroups and STR haplotypes in an Algerian population sample." International journal of legal medicine 122.3 (2008): 251-255.

Rosa, A., Ornelas, C., Jobling, M.A., Brehm, A., Villems, R. (2007). Y-chromosomal diversity in the population of Guinea-Bissau: a multi-ethnic perspective. BMC Evolutionary Biology, 10.1186(1471-2148), 7-124.

Rosa, Alexandra, et al. "Population data on 11 Y-chromosome STRs from Guine-Bissau." Forensic science international 157.2 (2006): 210-217.

Sezgin, E., Drosdak , A., McIntosh , C., Kessing, B., Lautenberger, J., Goedert, J. J., Phair, J. P., Troyer, J. L., Smith, M. W., Brien, S. J. O. (2010). Examination of disease based selection, demographic history and population structure in European Y chromosome haplogroup I. Institutes Health of National Henn,10.1038, 613–620.

Sher A, Seyed E H., Molecular Dissection of the Human Y-Chromosome. Gene 2002:283:1-10.

Singer M., (1982). Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. 76:67-112.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics, 139, 457–462.

Soares-Vieira, José A., et al. "Population and mutation analysis of Y-STR loci in a sample from the city of São Paulo (Brazil)." Genetics and Molecular Biology 31.3 (2008): 651-656.

Strachan T., Read A. Human Molecular Genetics. AMA Graphics Ltd. Preston, UK. 1997.

Stryer, Lubert,. Biochemisrty. Standford University. 1st ed., 1999.

Takezaki, Naoko, and Masatoshi Nei. "Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA." Genetics 144.1 (1996): 389-399.

Tishkoff SA, Kidd K K.(2004) Implications of biogeography of human populations for race and medicine. Nature 36:21-27.

Turrina, Stefania, Renzo Atzei, and Domenico De Leo. "Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay." International journal of legal medicine 120.1 (2006): 56-59.

Underhill, Peter A., and Toomas Kivisild. "Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations." Annu. Rev. Genet. 41 (2007): 539-564.

Urquhart A., Kimpton C., Downs T., and Gill P., (1994). Variation in short tandem repeats sequences- a survey of twelve microsatellite loci for use as forensic identification markers. Int. J. Leg. Med. 107, 13-20.

Valdes, Ana Maria, Montgomery Slatkin, and N. B. Freimer. "Allele frequencies at microsatellite loci: the stepwise mutation model revisited." Genetics 133.3 (1993): 737-749.

Willard H., and Waye J., (1987). Hierarchical order in chromosome-specific human alpha satellite DNA. Trends in Genet. 3, 192-198.

William B. Provine (1978). "The role of mathematical population geneticists in the evolutionary synthesis of the 1930s and 1940s". Studies of the History of Biology 1: 167–192.

Yang T., Hansen S., Oishi K., Ryder O., and Hamkalo A. (1982). Characterization of a cloned repetitive DNA sequence concentrated on the human X chromosome. Proc. Natl. Acad. Sci. USA 79, 6593-6587.

Yunis, Juan J., et al. "Geno-geographic origin of Y-specific STR haplotypes in a sample of Caucasian-Mestizo and African-descent male individuals from Colombia." Biomedica: revista del Instituto Nacional de Salud 33.3 (2013): 459-467.

Zheng, Lihong, et al. "Y chromosomal STR polymorphism in northern Chinese populations." Biological research 42.4 (2009): 497-504.

Zinck, H. H, Musharoff, S., Salit, J., Al-Ali, K. A., Chouchane, L., Gohar, A., Matthews, R., Butler, M.W., Fuller, J., Hackett, N. R., Crystal, R.G., Clark, A.G. (2010). Population Genetic Structure of the People of Qatar. American Journal of Human Genetics, 87, 17–25.

تشير الدراسة الى أن القدرة التمييزية بين الأفراد تصل الى ما يقارب 90% تقريبا بينما يصل تنوع ال (Haplotype) الى 99.885% تقريبا. و هو أمر هام في فهم درجة عدم التجانس و التناسق في التركيبة السكانية في دولة الامارات العربية المتحدة و يمكن أن ينعكس ذلك بالتالي على عوامل عدة منها نمط هجرة السكان من و الى المنطقة. التأثير الجغرافي. و التأثيرات الثقافية التي يمكن حدوثها تبعا لذلك. ثانيا: ان الدراسة توفر قاعدة عريضة من المعلومات المتعلقة بدراسة ال (Haplotype) و تخدم كقاعدة بيانات يمكن استخدامها في مجال الأدلة الجنائية. و تشير الدراسة الى أن القدرة التمييزية بين الأفراد تصل الى ما يقارب 90% تقريبا بينما يصل تنوع ال (Haplotype) الى 99.885% تقريبا. و هو أمر هام في فهم درجة عدم التجانس و التناسق في التركيبة السكانية في دولة الامارات العربية المتحدة و يمكن أن ينعكس ذلك بالتالي على عوامل عدة منها نمط هجرة السكان من و الى المنطقة. التأثير الجغرافي. و التأثيرات الثقافية التي يمكن حدوثها تبعا لذلك. ثانيا: ان الدراسة توفر قاعدة عريضة من المعلومات المتعلقة بدراسة ال (Haplotype) و تخدم كقاعدة بيانات يمكن استخدامها في مجال الأدلة الجنائية.

و ختاما، التركيبة السكانية في دولة الأمارات العربية المتحدة تعتبر تركيبة متنوعة و تماثل جينيا الدول المجاورة لها. و لا يظهر تحليل التباين الجزيئي أو ما يعرف ب (AMOVA) أية اختلافات جينية كبيرة يمكن ملاحظتها على التركيبة السكانية داخل دولة الامارات العربية المتحدة خصوصا أو على السكان القاطنين في دول الخليج العربي عموما.

الخلاصة

تستخدم تطبيقات (Y chromosome) و التي تعتبر خاصة بالذكور فقط و التي تحتوي على أطول جزء من الجين البشري الغير قابل للجمع و التوحيد, و يستخدم في الدراسات الجينية في مجال التركيبة السكانية.

تهدف الدراسة التي الى انشاء قاعدة معلوماتية فيما يتعلق بالكروموسوم الذكري لمواطني دولة الإمارات العربية المتحدة, مع الوضع فى الاعتبار, بأن دراستنا تعتبر الأكبر في هذا المجال من حيث أعداد الفئة التي أجري عليها الفحص. تم اجراء الفحوص الجينية من خلال استخدام 17 عامل وراثي عن طريق تقنية (STR'S و قد تم جمع و استخلاص البصمة الوراثية من 345 عينة من المتطوعين الذكور من مواطني الدولة من غير الأقارب.

أوضحت النسب من العينات التي تمت دراستها و تحليلها من سكان الدولة, بأن كل (Locus) و الذي هو عبارة عن جين على موقع معين من الكروموسوم, يملك (Allele) و يعتبر هو السائد, و يتوضح لنا أيضا أن (Allele) لمعظم المواضع (loci) تتجمع على نطاق ضيق بحيث يكون ما نسبته من 60% - 80% من السكان تتقاسم أليل محدد لهذه المواقع الجينية على الكروموسوم.

وقد تلاحظ بأن أعلى تنوع جيني في موقع DYS458 و موقع DYS385-A و موقع-DYS385 B و يبلغ 0.9 وبالتالي، يجب اعتبارها على أساس أنها أكبر متغير و الأفضل من حيث قدر المعلومات التي يمكن استخدامها في فحوص الأدلة الجنائية. بينما ال (Loci) ذات التنوع الأقل يعتبر مصدرا قليل المعلومات و المثال عليه في الموقع DYS392 و حصلنا علي تنوع جيني و قدره 0.437.

يعتبر مجتمع دولة الإمارات العربية المتحدة مجتمعا مختلطا حيث تم العثور على 301 (Haplotype) منهم 271 وجدت عيناتهم فريدة في المجتمع بينما وجدت 22 منهم منتشرة بين اثنين من السكان. و وجدت ثلاث حالات لأفراد يتشاركون في اربعة, خمسة و ستة من نفس ال (Haplotype). بالإضافة الى وجود ثلاثة من ال (Haplotypes) المختلفة تتشارك بواسطة خمسة من الأفراد. و يمكن ارجاع ذلك الى ان المجتمع يتشارك في نفس السلالة السكانية.

جامعة الإمارات العربية المتحدة
كلية العلوم
قسم علوم الحياة

دراسة التنوع الوراثي للكروموزوم الذكري (Y-chromosome) في مجتمع الامارات العربية المتحدة

رسالة مقدمة من /

صفاء يعقوب يوسف الخياط الحمادي

مقدمة إلى /

جامعة الإمارات العربية المتحدة

استكمالا لمتطلبات الحصول على درجة الماجستير في علوم البيئة

تحت اشراف الدكتور خالد أميري و الدكتور عبد المجيد الخاجه والدكتور أحمد المرزوقي

أغسطس 2014